计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2020,29(12):216−221 [doi: 10.15888/j.cnki.csa.007684]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于快速边界攻击的黑盒对抗样本生成方法①

郭书杰

(大连东软信息学院 智能与电子工程学院, 大连 116023)
通讯作者: 郭书杰, E-mail: guoshujie@neusoft.edu.cn

摘　要: 深度学习技术在不同领域有着广泛的应用, 然而一个训练好的深度学习模型很容易受到干扰而得出错误的结果, 从而引发严重的安全问题. 为了检验深度学习模型的抗干扰性, 提高模型的安全性和鲁棒性, 有必要使用对抗样本进行对抗评估和对抗训练. 有目标的黑盒对抗样本的生成方法具有较好的实用性, 是该领域的研究热点之一. 有目标的黑盒对抗样本生成的难点在于, 如何在保证攻击成功率的前提下提高对抗样本的生成效率. 为了解决这一难点, 本文提出了一种基于快速边界攻击的有目标攻击样本生成方法. 该方法包括线上的搜索和面上的搜索两步. 线上的搜索由单侧折半法来完成, 用于提高搜索效率; 面上的搜索通过自适应调节搜索半径的随机搜索完成, 用于提高搜索的广度. 通过对 5 组图片的实验结果验证了方法的可行性.

关键词: 黑盒攻击; 对抗样本; 深度学习

引用格式: 郭书杰.基于快速边界攻击的黑盒对抗样本生成方法.计算机系统应用,2020,29(12):216–221. http://www.c-s-a.org.cn/1003-3254/7684.html

Black Box Adversarial Examples Generation Method Based on Fast Boundary Attack

GUO Shu-Jie

(School of Intelligence and Electronic Engineering, Dalian Neusoft University of Information, Dalian 116023, China)

Abstract: Deep learning is widely used in different fields. However, a well-trained deep learning model may be easily disturbed and gives wrong results, which causes serious safety problems. In order to test the robustness of deep learning model, researchers attack the model by all kinds of adversarial examples. The generation method of black box adversarial examples with targets, which has sound practicability, becomes a hot issue. The difficulty of black box adversarial examples generation lies in how to improve the generation efficiency under the premise of the success rate of the attack. In order to solve this difficulty, this study proposes a new method of target adversarial sample generation based on fast boundary attack. This method includes two steps: sampling along the line and sampling on the sphere. The first step is completed by the one side half search to improve search efficiency. The second step is completed by random search with adaptive adjustment of search radius, which is used to improve the search scope. The feasibility of the algorithm is verified by experimental results of five groups of pictures.

Key words: black box attack; adversarial examples; deep leaning

基于神经网络的深度学习技术已经被成功地应用于计算机视觉[1,2]、语音识别[3]和自然语言处理[4−9]等多个领域. 特别是在机器视觉中的图像识别方面, 深度学习技术取得了非常大的成就. 尽管如此, 深度学习技术自身也存在着比较严重的安全问题. Szegedy 等[10]发现在使用深度学习技术进行图像识别时, 只要改动图片上的一个像素, 就能让神经网络识别错误, 甚至还可以诱导它返回特定的结果. 在自动驾驶、人脸识别、

① 收稿时间: 2020-04-09; 修改时间: 2020-05-10; 采用时间: 2020-05-18; csa 在线出版时间: 2020-11-30

语音识别、CT 影像分类等典型的深度学习应用中, 错误的识别结果将会带来非常严重的后果. 因此很多研究者开始关注深度学习模型的抗干扰能力的问题.

为了检验深度学习模型的抗干扰性和鲁棒性, 研究人员提出了对抗样本的概念. 所谓对抗样本就是在一个已经正确分类的样本中, 添加细微干扰形成的新样本, 该样本可以使训练好的模型以较高的置信度给出错误的分类结果[11]. 国内外研究者提出了多种对抗样本生成方法[12–19]. 按照不同的规则, 可以将这些方法划分成不同种类. 按照其生成方式和原理的不同, 可以分为部分像素添加扰动和全像素添加扰动两类. 按照生成过程是否需要知道模型内部结构与参数, 可以分为白盒方法和黑盒方法. 需要知道模型内部机构与参数的生成方法叫白盒方法, 反之叫黑盒方法. 根据对抗规则的不同又可以分为有目标对抗和无目标对抗. 有目标对抗是指对抗样本需要使模型给出某种指定的错误类别; 无目标对抗则只要求模型给出错误分类结果即可. Su 等提出了一种黑盒对抗样生成方法 ONE-PIXEL[20], 该方法将对抗样本的生成过程转换为一个条件优化问题, 然后使用差分进化算了来求解该问题, 并最终得到对抗样本. 该方法可以对梯度难以计算和不可微的网络进行攻击, 具有良好的灵活性. 然而, 由于只改变了原始图像的一个像素, 该方法的攻击成功率相对较低, 特别是有目标攻击的成功率. Dong 等在借鉴 I-FGSM 和 ILCM 方法的基础上, 提出了 MI-FGSM 黑盒攻击方法[21]. 该方法通过将动量迭代来替换梯度迭迭代, 使得在迭代过程具有更加稳定的更新方向, 从而降低陷入局部最优的概率. 虽然该方法对添加的噪声方向进行了平滑, 但是随着迭代次数增加, 边界效应依然存在. 为了解决这一问题, Shi 等提出了 Curls & Whey 方法[19]. Curls & Whey 方法通过使迭代轨迹的多样化和压缩噪声的幅度来提高生成的对抗样本的质量. 由于 MI-FGSM 和 Curls & Whey 均为基于迁移的攻击, 所以他们均不能保证个体级别的攻击成功. Brendel 等提出了一种基于决策的有目标黑盒对抗样本生成方法[15], 该方法能够保证攻击的成功率, 但需要较多的模型访问次数, 因此效率相对较低. 有目标的黑盒攻击的难点在于, 如何在保证攻击成功率的前提下提高对抗样本的生成效率. 为了解决这一难点, 本文提出一种应用于图像分类领域的全像素添加扰动的黑盒对抗方法, 该方法主要针对有目标对抗, 同时也适用于无目标对抗样本的生成.

1 基于快速边界攻击的黑盒对抗样本生成方法

1.1 面向图像分类的有目标黑盒攻击

深度神经网络可以完成各种不同的分类任务, 本文讨论的是图像分类任务中的深度神经网络模型对抗样本的生成方法. 在用于图像分类的神经神经网络中, 图片的每个通道通常用矩阵 A_{nn} 表示. 其中 n 表示图像的行数和列数, 每个元素取 0–255 之间的整数. 对于一个深度学习模型 M, 要对一张正确分类为 N 的图片 X 生成一个干扰目标为 L 的黑盒攻击样本, 就是在 X 上添加较少的噪声干扰得到样本 X', 使得 M 对 X'的分类结果为 L. 也就是:

$$\begin{cases}\min\|\rho\|_2 \text{ s.t. } M(x+\rho)=L\\ x+\rho\in[0-255]^n,\ \rho\to 0\end{cases} \tag{1}$$

式 (1) 中, ρ 是需要加入的干扰噪声.

1.2 相关定义

为了便于问题描述, 给出以下定义.

定义 1. 决策空间: 在一个图像分类神经网络中, 所有被分类为 A 的图像组成的集合, 就叫做 A 的决策空间 S_A, 也就是 A 的决策空间:

$$S_A=\{P_i|pre(P_i)=A\} \tag{2}$$

定义 2. 决策边界: 在一个图像分类神经网络中, 分类 A 的决策边界是指 A 的决策空间的最外层, 也就是那些即便做极其微小的改动都会改变其分类结果的图像的集合. A 的决策边界:

$$\begin{cases}B_A=\{P_i|pre(P_i)=A\wedge pre(P_i+\delta)\neq A)\}\\ \delta\to 0\end{cases} \tag{3}$$

定义 3. 图像间的距离: 本文中使用欧氏距离来定义两张图片间的距离. 令图片 P_1 的矩阵为 X, 图片 P_2 的矩阵为 Y, 则 P_1 和 P_2 之间的距离为:

$$D(P_1,P_2)=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2} \tag{4}$$

基于上述定义, 对于一张分类为 M 的图片 P_m, 要生成一个分类为 N 的对抗样本, 也就是要在 N 的决策空间中找到一个点 P_n, 使得 P_m 和 P_n 的距离尽可能小. 即: $P_n=\min\{D(P_m,P_n)|P_n\in B_N\}$

根据决策边界的定义不难看出, 最理想的对抗样本一定在 N 的决策边界 B_N 上, 如图 1 所示.

1.3 快速边界攻击法

边界攻击就是沿着某一分类 N 的临近决策边界 BN 寻找距离被攻击目标最近的点的过程. 如图 2 所示.

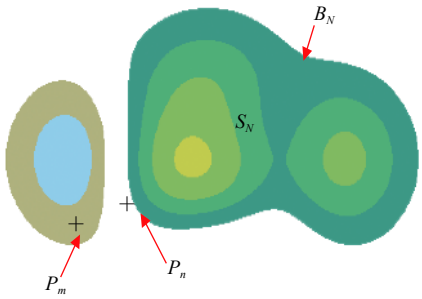

图 1 对抗样本示例

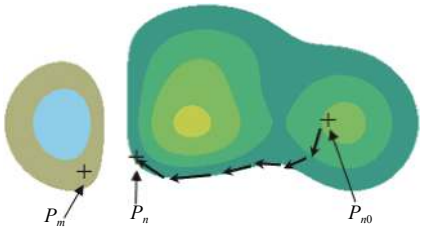

图 2 边界攻击过程

为了能够快速找到最优攻击样本 P_n, 设计了一种快速边界攻击法. 快速边界攻击首先通过单侧折半法快速找到干扰样本和被攻击样本的近似边界所在, 然后再以可自动调节的步长沿着决策边界 B_N 探索, 直至找到满足停止条件的对抗样本. 具体的步骤如下.

第 1 步. 初始化攻击样本. 从决策空间 S_N 中随机选出一张图片 P_{n0} 作为初始攻击样本 P_{adver}.

第 2 步. 找到 P_m 与 P_{adver} 之间的近似边界点. 使用单侧折半查找法在 P_m 与 P_{adver} 之间的连线上找到距离决策边界 BN 相对较近的点 $P_{boundary}$, 并将该点赋值给 P_{adver}. 单侧折半查找法的具体步骤如下.

① 首先根据图像间的距离公式, 确定被攻击目标 P_m 与攻击样本 P_{adver} 之间的中点 $P_{mid}=(P_m+P_{adver})/2$.

② 使用神经网络对 P_{mid} 进行分类预测, 得到分类结果 C_{mid}; 若 $C_{mid}=N$, 则将 P_{mid} 赋值给 P_{adver}; 若 $C_{mid}\neq N$, 则在后 (右) 半个区域 P_{mid} 和 P_{adver} 之间继续进行折半查找, 直至找到分类结果为 N 的 P_{mid}, 将 P_{mid} 赋值给 P_{adver}. 单侧折半法的具体过程如图 3 所示.

第 3 步. 沿着 N 的近似边界随机探索更优样本. 以自适应步长 δ 在 P_{adver} 附近随机寻找 n 个对抗样本, 将这些样本中距离 P_m 最近的分类结果为 N 的点赋值给 P_{adver} 并转到第 2 步继续运行, 直至找到满足停止条件的对抗样本. δ 的大小决定了算法在 P_{adver} 附近的搜索半径, 当 δ 比较小时, 算法只能在 P_{adver} 较近的区域搜索, 由于搜索到的点大多与目标点 P_n 较远, 所以搜索效率不高; 当 δ 比较大时, 算法的搜索范围可能会超过决策边界 B_N, 从而使得无法找到满足条件的样本, 导致搜索停滞. 为了在提高算法的搜索效率, 步长 δ 的初始值取 0.1, 随着算法的进行, 自动调节 δ 的值, 其调节策略如下. 使用神经网络对以 δ 为步长在 P_{adver} 附近随机寻找 n 个对抗样本进行预测, 计算预测结果中分类 N 的平均值 MSN. 该平均值越大, 说明 n 个对抗样本中属于决策空间 S_N 的样本越多, 距离决策边界 B_N 越远. 为了提高优化效率, 需要让 δ 增大. 相反, 该平均值越小, 说明步长 δ 设置得过大, 使得 n 个对抗样本中较多的样本已经越过了策边界 B_N, 需要减小 δ 的值. 为了确定自适应调节参数, 对调节时机 (即 MSN 的值取多少时进行调节)、调节量 (即 δ 值的缩放系数) 进行了对比实验. 实验以达到 0.9 的样本优化率 (见定义 5) 所需的模型访问次数为标准来评价算法的搜索效率, 从而确定参数的优劣. 实验结果显示, 当 MSN 的值介于 0.3–0.7 时, 算法能够保持相对稳定的搜索效率. 依据实验结果, 采用如下调节方案: 当 MSN 的值大于 0.7 时 δ 扩大为原来的 1.1 倍; 当 MSN 的值大于 0.8 时 δ 扩大为原来的 1.3 倍; 当 MSN 的值大于 0.9 时 δ 扩大为原来的 1.7 倍. 当 MSN 的值小于 0.3 时 δ 缩小为原来的 0.9 倍; 当 MSN 的值小于 0.2 时 δ 缩小为原来的 0.7 倍; 当 MSN 的值小于 0.1 时 δ 缩小为原来的 0.5 倍. 快速边界攻击法的算法如算法 1.

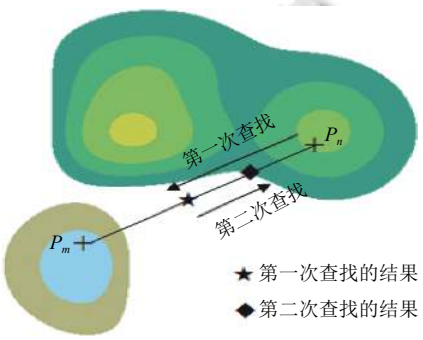

图 3 单侧折半法的查找过程示例

算法 1. 快速边界攻击样本生成算法

输入: 被攻击原始图片 P_m, 错误分类 N, 待攻击的分类模型 CNN-M.
输出: 攻击样本 P_n.

1) 初始化相关参数;
2) 从 N 的决策空间中随机选出一张图片作为初始攻击样本 P_{adver};
3) **while** (true)
4) 使用单侧折半查找法查找 P_m 和 P_{adver} 之间的临近边界点, 并将其赋值给 P_{adver};
5) 以 δ 为步长在 P_{adver} 附近随机生成 n 个样本 $PR_n=\{P_1, P_2, \cdots, P_n\}$;
6) 使用分类模型 CNN-M 对 PR_n 中的每个样本 P_i 进行分类预测;

7) 将这 n 个样本 PR_n 中分类预测结果为 N 且距离 P_m 最近的点赋值给 P_{adver};
8) 根据对这 n 个样本的分类预测结果调节步长 δ 的值;
9) **if** (满足终止条件)
10) **break**;

一张图片的决策空间是立体的多维球面, 为了能够高效地在其决策边界附近找到一个近似最优解, 快速边界攻击法的搜索过程分为线上的搜索和面上的搜索两步. 线上的搜索由单侧折半法来完成, 用于快速找到在 P_{n0} 到 P_m 的连线上距离决策边界较近的点 P_{adver}; 面上的搜索通过以自适应步长 δ 在 P_{adver} 附近的随机搜索来完成. 通过快速的线上搜索来提高算法的速度, 使用面上的搜索来提高搜索的广度, 这两个步骤的结合, 既保证了算法具有良好的搜索效率, 又保证了算法搜索结果的质量. 由于该方法采用的是从错误分类 N 的决策空间逐渐向正确分类 M 的决策空间靠近的搜索方法, 而且在搜索过程中始终保证每一轮搜索的最优结果均在 N 的决策空间中, 所以可以保证最终生成的对抗样本具有可靠的攻击成功率.

使用该方法生成一张图片的有目标对抗样本的过程如图 4 所示. 假设要生成一张分类结果为“Siamese_cat”的图片 A 的对抗样本, 使得深度学习模型将其误分为“Labrador_retriever”. 在开始时, 从分类为“Labrador_retriever”的样本中随机选择一张图片 B, 然后使用快速边界攻击法在 B 的近似决策边界上寻找距离 A 最近的图片作为 A 的对抗样本 B'. 不难看出, 随着寻优过程的推进 B'与 A 的距离逐渐减少, 攻击样本与原始图像 A 的差别也越来越小.

注：图片下方的数字是该图片与图片 A 之间的距离

图 4 快速边界攻击法的具体示例

1.4 无目标对抗样本生成方案

快速边界攻击法主要适用于有目标对抗, 但它也能够实现无目标攻击, 只需改变初始化攻击样本的生成方法即可. 在有目标对抗样本生成时, 初始攻击样本是从目标类决策空间中随机选出一张图片. 为了提高算法的效率, 在使用该方法生成某一分类为 N 的图片 A 的无目标对抗样本生成时, 首先从其他非 N 分类的决策空间中随机选出 m 个样本, 然后分别计算这 m 个样本与图片 A 的距离, 从中选出距离最小的样本作为初始攻击样本. 也就是说, 在进行无目标对抗样本生成时, 选择 m 个随机样本中与 A 的相似度最高的图片作为初始攻击样本, 以便提高对抗样本的生成效率.

2 实验及结论

2.1 可行性实验

为了检验方法的可行性, 在 Windows 10 平台上使用编程实现了快速边界攻击法, 并进行了 5 组图片的样本生成实验. 实验中对抗的网络模型为 ResNet50, 使用的测试数据如图 5 所示. 其中上面的图像为初始对抗样本, 下边的图像为被攻击目标样本, 从左至右分别称为 a 组、b 组、c 组、d 组和 e 组.

在评价算法的效率时, 需要选择合适的参数作为对比对象, 为了便于描述该参数, 给出以下定义.

定义 4. 模型访问次数: 在算法运行过程中, 调用深度学习模型进行分类预测的总次数.

定义 5. 样本优化率: 令给定的初始对抗样本为 A, 被攻击目标样本为 B, 某代优化得到的临时对抗样本为 A' 则样本优化率 R_o 的定义如下:

$$R_o = 1 - mean\left(D\left(A,A'\right)/D(A,B)\right) \tag{5}$$

其中, $D\left(A,A'\right)$表示与 A'与 A 的距离, $D(A,B)$表示与 B 与 A 的距离, $mean$ 表示取平均数. 由于$D\left(A,A'\right)$和$D(A,B)$一般为包含 3 个元素的向量, 所以需要对该向量中的元素取平均数.

图 5 对比实验用的图片

基于上述定义, 给定优化率时的模型访问次数可以表示算法的对抗样本生成效率的大小, 对于相同的样本优化率来说, 模型访问次数越少, 算法的效率就越高. 五组实验数据的实验结果如图 6 所示. 由实验结果不难看出, 对于 5 组图片, 算法均能在约 4400 次模型访问后, 达到 0.8 的模型优化率; 在 12 000 左右次模型访问后, 达到 0.9 的模型优化率. 算法在初期 (样本优化率<0.6 时) 生成效率相差不大; 但在中后期会有不同的表现, e 组最快, a 组最慢.

2.2 效率实验

为了检验方法的生成效率, 与 Wieland Brendel 的 Boundary 方法[15] 做了对比实验. 实验的对抗的网络模型、测试数据、算法效率评价方法与 2.1 节中可行性实验的相同. 实验结果如图 7 所示. 对于 5 组图片, 快速边界攻击法的效率较 Boundary 方法均有不同程度的提高, 能够用相对较少的模型访问次数来达到相同的本优化率.

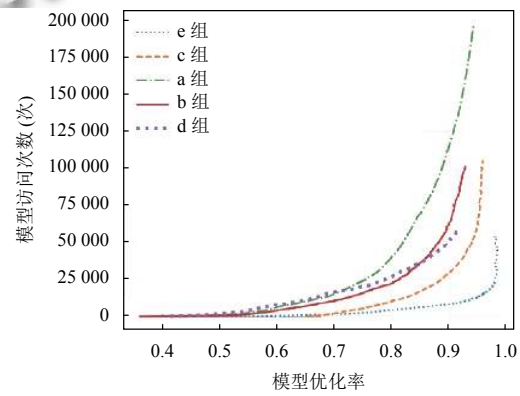

图 6 可行性实验结果

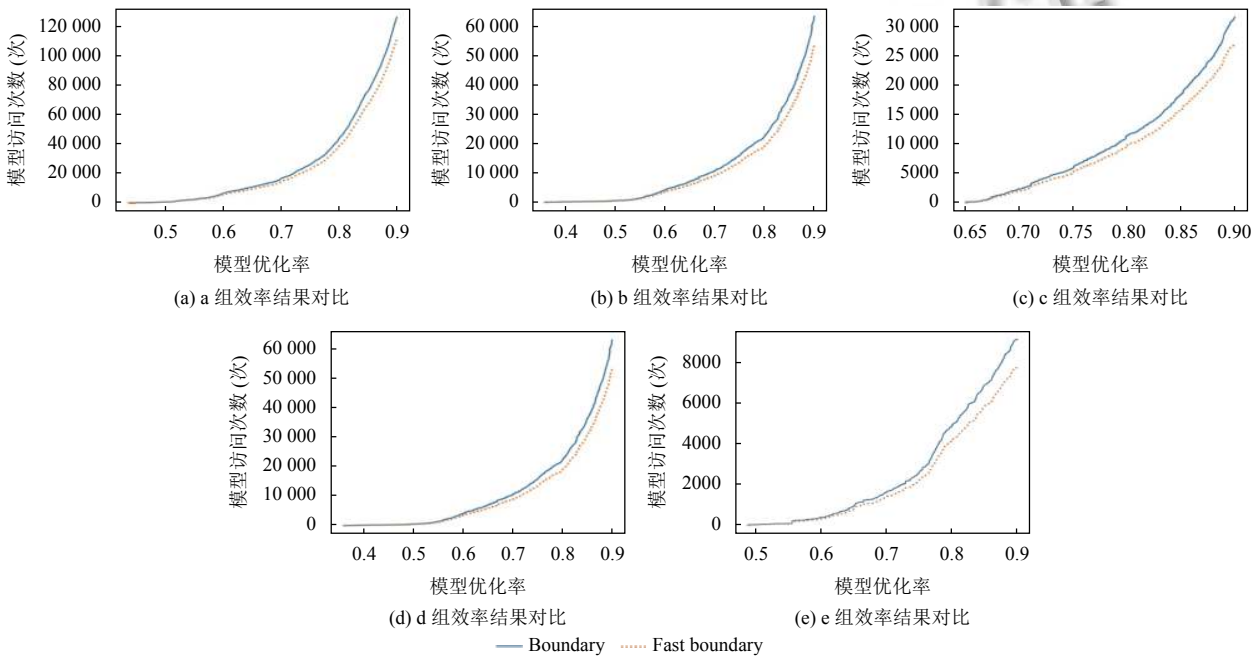

图 7 效率实验结果

2.3 结论

快速边界攻击法较为简单, 易于实现, 具有较好的通用性; 能够生成无目标对抗样本和有目标对抗样本, 而且属于比较有应用价值的黑盒对抗样例生成方法; 与 Boundary 方法相比, 快速边界攻击法具有相对较好的生成效率. 但由于每个分类的决策空间相对较大, 为了找到近似最小扰动, 该方法的访问次数还是比较大, 所以生成过程比较耗时, 因此不适用于对实时性有要求的对抗样本的生成.

参考文献

1 Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Proceedings of the 25th International Conference on Neural Information Processing Systems. Lake Tahoe, NV, USA. 2012. 1097−1105.

2 Taigman Y, Yang M, Ranzato M, *et al*. DeepFace: Closing the gap to human-level performance in face verification. Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA. 2014. 1701−1708.

3 Dahl GE, Yu D, Deng L, *et al*. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. IEEE Transactions on Audio, Speech, and Language Processing, 2012, 20(1): 30−42.

4 Collobert R, Weston J. A unified architecture for natural language processing: Deep neural networks with multitask learning. Proceedings of the 25th International Conference on Machine Learning. Helsinki, Finland. 2008. 160−167.

5 Zhang X, Zhao JB, LeCun Y. Character-level convolutional networks for text classification. Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal, QC, Canada. 2015. 649−657.

6 Kim Y, Jernite Y, Sontag D, *et al*. Character-aware neural language models. Proceedings of the 30th AAAI Conference on Artificial Intelligence. Phoenix, AZ, USA. 2016. 2741−2749.

7 Pang B, Lee L, Vaithyanathan S. Thumbs up? Sentiment classification using machine learning techniques. Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing. Philadelphia, PA, USA. 2002. 79–86.

8 Sutskever I, Vinyals O, Le QV. Sequence to sequence learning with neural networks. Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal, QC, Canada. 2014. 3104−3112.

9 Maas AL, Daly RE, Pham PT, *et al*. Learning word vectors for sentiment analysis. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Portland, OR, USA. 2011. 142–150.

10 Szegedy C, Zaremba W, Sutskever I, *et al*. Intriguing properties of neural networks. Proceedings of the 2nd International Conference on Learning Representations. Banff, AB, Canada. 2014.

11 Akhtar N, Mian A. Threat of adversarial attacks on deep learning in computer vision: A survey. IEEE Access, 2018, 6: 14410–14430. [doi: 10.1109/ACCESS.2018.2807385]

12 王文琦, 汪润, 王丽娜, 等. 面向中文文本倾向性分类的对抗样本生成方法. 软件学报, 2019, 30(8): 2415−2427. [doi: 10.13328/j.cnki.jos.005765]

13 张思思, 左信, 刘建伟. 深度学习中的对抗样本问题. 计算机学报, 2019, 42(8): 1886–1904.

14 潘文雯, 王新宇, 宋明黎, 等. 对抗样本生成技术综述. 软件学报, 2020, 31(1): 67–81. [doi: 10.13328/j.cnki.jos.005884]

15 Brendel W, Rauber J, Bethge M. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. Proceedings of the 6th International Conference on Learning Representations. Vancouver, BC, Canada. 2018.

16 Cisse M, Adi Y, Neverova N, *et al*. Houdini: Fooling deep structured visual and speech recognition models with adversarial examples. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, CA, USA. 2017. 6977–6987.

17 Co KT, Muñoz-González L, de Maupeou S, *et al*. Procedural noise adversarial examples for black-box attacks on deep convolutional networks. Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security. London, UK. 2019. 275–289.

18 Chen PY, Zhang H, Sharma Y, *et al*. ZOO: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security. Dallas, TX, USA. 2017. 15–26.

19 Shi YC, Wang SY, Han YH. Curls & Whey: Boosting black-box adversarial attacks. Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, CA, USA. 2019. 6512–6520.

20 Su JW, Vargas VV, Sakurai K. One pixel attack for fooling deep neural networks. IEEE Transactions on Evolutionary Computation, 2019, 23(5): 828–841. [doi: 10.1109/TEVC.2019.2890858]

21 Dong YP, Liao FZ, Pang TY, *et al*. Boosting adversarial attacks with momentum. Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA. 2018. 9185–9193.