北京航空航天大学学报
Journal of Beijing University of Aeronautics and Astronautics
ISSN 1001-5965,CN 11-2625/V

《北京航空航天大学学报》网络首发论文

题目：一种基于攻击距离的对抗样本攻击组筛选方法
作者：刘洪毅，方宇彤，文伟平
DOI：10.13700/j.bh.1001-5965.2020.0529
收稿日期：2020-09-21
网络首发日期：2021-01-20
引用格式：刘洪毅，方宇彤，文伟平．一种基于攻击距离的对抗样本攻击组筛选方法．北京航空航天大学学报. https://doi.org/10.13700/j.bh.1001-5965.2020.0529

网络首发：在编辑部工作流程中，稿件从录用到出版要经历录用定稿、排版定稿、整期汇编定稿等阶段。录用定稿指内容已经确定，且通过同行评议、主编终审同意刊用的稿件。排版定稿指录用定稿按照期刊特定版式（包括网络呈现版式）排版后的稿件，可暂不确定出版年、卷、期和页码。整期汇编定稿指出版年、卷、期、页码均已确定的印刷或数字出版的整期汇编稿件。录用定稿网络首发稿件内容必须符合《出版管理条例》和《期刊出版管理规定》的有关规定；学术研究成果具有创新性、科学性和先进性，符合编辑部对刊文的录用要求，不存在学术不端行为及其他侵权行为；稿件内容应基本符合国家有关书刊编辑、出版的技术标准，正确使用和统一规范语言文字、符号、数字、外文字母、法定计量单位及地图标注等。为确保录用定稿网络首发的严肃性，录用定稿一经发布，不得修改论文题目、作者、机构名称和学术内容，只可基于编辑规范进行少量文字的修改。

出版确认：纸质期刊编辑部通过与《中国学术期刊（光盘版）》电子杂志社有限公司签约，在《中国学术期刊（网络版）》出版传播平台上创办与纸质期刊内容一致的网络版，以单篇或整期出版形式，在印刷出版之前刊发论文的录用定稿、排版定稿、整期汇编定稿。因为《中国学术期刊（网络版）》是国家新闻出版广电总局批准的网络连续型出版物（ISSN 2096-4188，CN 11-6037/Z），所以签约期刊的网络版上网络首发论文视为正式出版。

一种基于攻击距离的对抗样本攻击组筛选方法

刘洪毅[1]，方宇彤[1]，文伟平[1]✉

（1. 北京大学软件与微电子学院，北京 102600）
*通信作者 E-mail： weipingwen@ss.pku.edu.cn

摘要 黑盒对抗样本生成过程中通常会指定一个攻击组，包括一个原始样本和一个目标样本，使得生成的对抗样本与原始样本差别不大，但其分类与目标样本一致。针对攻击组的攻击难度不同导致的攻击不稳定的问题，以图像识别领域为例，首先设计了基于决策边界长度的攻击距离度量方法，为攻击组的攻击难易程度提供了度量方法。并在此基础上设计了基于攻击距离的对抗样本攻击组筛选方法，在攻击开始前就筛去难以攻击的攻击组，从而实现在不修改攻击算法的前提下，提升攻击效果。实验表明，相比于筛选前的攻击组，筛选后的攻击组的总体效果提升 42.07%，攻击效率提升 24.99%，稳定方差 76.23%。建议所有利用攻击组的对抗样本生成方法在攻击前，先进行攻击组的筛选，以稳定并提高攻击效果。

关键词 对抗样本；黑盒；决策边界；筛选；图像识别
中图分类号 TP319 文献标识码：A DOI: 10.13700/j.bh.1001-5965.2020.0529

A method of filtering the attack pairs of adversarial examples based on attack distance

LIU Hongyi[1], FANG Yutong[1], WEN Weiping[1]✉

（1. School of Software and Microelectronics, Peking University, Beijing 102600, China）
*E-mail: weipingwen@ss.pku.edu.cn

Abstract During the generation of black-box adversarial examples, an attack pair is usually specified, including a source example and a target example. The purpose is to let the generated adversarial example only has little difference from the source example, but its classification is consistent with the target example. Aiming at the problem of the instability of adversarial attacks caused by different attack difficulty of attack pairs, taking the image recognition field as an example, firstly, this paper presented an attack distance measurement method based on the length of the decision boundary, which provided a measurement method for the attack difficulty of attack pairs. Then, this paper designed a filtering method based on attack distance of the attack pairs, which filtered out attack pairs that are difficult to attack before the attack starts, so this method can improve the attack effect without modifying the attack algorithm. Experiments show that compared with the attack pairs before filtering, the filtered attack pairs improve the overall attack performance by 42.07%, improve the attack efficiency by 24.99%, and stabilize the variance by 76.23%. It is recommended that all methods of generating adversarial examples using attack pairs should filter attack pairs before attack to stabilize and improve the attack effect.

Key words adversarial examples; black box; decision boundary; filtering; image recognition

随着深度学习算法及计算设备的快速发展，深度学习由于其优秀的效果与性能被广泛应用与各个领域，例如语音识别、图像识别、人脸识别、自动驾驶等。在计算机视觉领域，随着 2012 年 Alex Krizhevsk 等人设计了 AlexNet[1]赢得了 ImageNet LSVRC-2012 比赛的冠军，奠定了卷积神经网络在深

收稿日期：2020-09-21
基金项目：国家自然科学基金（基金号 U1736218）
作者简介：刘洪毅,男，硕士研究生。主要研究方向：软件安全、软件分析、智能安全。Email：hongyiliu@pku.edu.cn 方宇彤,女，硕士研究生。主要研究方向：软件安全、软件分析、智能安全。Email：fangyutong@pku.edu.cn 文伟平,男，博士，教授，博士生导师。主要研究方向：系统与网络安全、大数据与云安全及智能计算安全研究。Email：weipingwen@ss.pku.edu.cn

Fund： National Natural Science Foundation of China (U1736218)
网络首发时间： 2021-01-20 08:56:15 网络首发地址： https://kns.cnki.net/kcms/detail/11.2625.V.20210119.1721.001.html

度学习应用中的地位。在之后的发展中，学者们先后提出了 GoogLeNet[2]、VGG[3]、ResNet[4]、DenseNet[5]等深层网络模型，使得基于深度学习的图像识别技术成为了图像识别领域的主流，并使得识别精准度在 ImageNet 上达到了 95%以上。在带来便利的同时，深度学习本身也存在着一定的安全问题，这引起了安全领域的极大关注。例如，Szegedy[6]很快发现了深层神经网络在图像识别领域的一个有趣的弱点，即深度神经网络容易被对抗样本欺骗。这种攻击的表现形式是对图像添加一个微小的扰动，并且这些扰动对于人类视觉来讲，是几乎不可察觉的，但会使得分类器分类错误。

随着研究的深入，学者们将攻击方法进行了分类。根据攻击者掌握的知识，对抗攻击可分为白盒攻击和黑盒攻击。其中，白盒攻击是指攻击者完全了解模型的架构和参数。白盒模型在实际的威胁模型中并不常见，更为广泛的应用场景是黑盒模型。因此我们更关注黑盒环境下的对抗样本生成技术。黑盒攻击指攻击者无法获取模型全部信息，只能通过对模型的使用来观察输入输出并展开攻击。

黑盒攻击方法主要有两种，基于可转移性的对抗样本生成方法和基于查询的对抗样本生成方法。可转移性是对抗样本的一种性质，即针对某分类器模型生成的对抗样本也可能欺骗其他分类器模型。基于可转移性的对抗样本生成方法（transfer-based adversarial example generation，TAEG）主要利用对抗样本的可转移性，针对一个替代模型进行白盒攻击，例如基于梯度的对抗样本生成方法，包括 FGSM[7]、I-FGSM[8]、PGD[9]、MI-FGSM[10]、CW[11]、JSMA[12]、Deepfool[13]等，并使用产生的对抗样本攻击目标模型，例如 Substitute Model[14]、DI-2-FGSM[15]、Ensemble Attack[16]。基于可转移性的对抗攻击可以仅查询目标模型一次就完成攻击，但其攻击成功率较低，若非目标模型和替代模型十分相似，即便是目前鲁棒性最好的方法，攻击成功率也难以令人满意[15]。而基于查询的对抗样本生成方法（query-based adversarial example generation，QAEG）已经被实践[17, 18]证明可以应用在现实模型上。QAEG 仅需要对目标模型进行一定的查询，根据查询结果不断优化对抗样本，即可实现黑盒攻击，并且成功率一般都较高。

目前流行的 QAEG 方法都是基于攻击组的，例如 NES Attack[17]、CMA Attack[18, 19]、Boundary Attack[20]等。这些方法引入原始样本与目标样本（我们称这一对样本为一个攻击组）共同参与攻击，当对抗样本生成任务陷入局部最优解时，为对抗样本进化提供了基本的方向，即向着原始样本靠近，可以在一定程度上缓解由于估算梯度不准确而导致的对抗样本陷入局部最优的情况。但是，由于引入了攻击组来指导攻击，针对一个对抗样本生成任务，如何合理的选择对抗攻击组将成为一个挑战。由于对抗攻击的目的仅是使得目标模型错误分类。但在现有的研究中，由于引入了攻击组的指导，相当于加入了一个限制条件，当生成任务到达局部最优解时，只能向着原始样本靠近。但原始样本与目标样本可能并不容易实现攻击，最终导致攻击失败，或者所需查询次数过多，例如 CMA Attack 实现攻击的平均查询次数为 60,000 次，但最坏的情况可能达到 200,000 次[18]。

为了解决以上挑战，我们以图像识别领域为例，设计了一种基于决策边界长度的攻击距离度量方法，探测原始图片到目标分类决策边界的垂直距离，并计算需要进化的决策边界长度，以度量两图片的攻击距离，并为该攻击组的攻击难易程度提供参考。同时，设计了基于攻击距离的对抗样本攻击组筛选方法，首先利图片间的攻击距离，选取较为容易完成攻击的攻击组，然后再展开对抗样本生成任务。最后，我们通过实验验证，筛选后的攻击组，对于多个攻击方法有效，提升总体攻击效果 42.07%，提升攻击效率 24.99%，稳定方差 76.23%。

综上所述，我们主要有以下贡献：

设计了一种基于决策边界长度的攻击距离度量方法，首次为图片间对抗攻击的难易程度提供了度量的方法，为攻击组的选取提供了参考。

设计了一种基于攻击距离的对抗样本攻击组筛选方法，在不改变攻击算法本身的情况下，降低了完成攻击所需的平均查询次数，提升总体攻击效果 42.07%，提升攻击效率 24.99%，稳定方差 76.23%。

1　相关工作

1.1 深度学习

深度学习[21]是机器学习领域的一个新的研究方向，它目前正以前所未有的规模被用于破解各种复杂的科学难题。例如，深度神经网络(Deep Neural Networks，DNN)在图像识别、重建脑回路[22]、DNA 突变分析[23]等各种任务上都取得了显著的成功。DNN 也成为语音识别[24]、自然语言理解[25]、无人机[26]、机器人[27]、和人脸识别 ATM[28]中许多具有挑战性任务的首选解决方案。显然，深度学习解决方案，尤其是那些源自计算机视觉问题的解决方案，将在我们的日常生活中发挥重要作用。

1.2 对抗样本

对于深度学习分类器来说，对抗样本是一种特殊的样本，其与原始样本的差别不大，但却可以使得分类器分类错误。对抗样本生成的基本思想是对于原始图片 x，设计一个尽量小的扰动 δ，能够使得分类器 C 分类错误，如公式（1）所示，上述场景被称为非目标攻击，即只要使得分类器分类错误即可。

$$\min \| \delta \| \ s.t. \ C(x+\delta) \neq C(x) \tag{1}$$

同时，还有一种更强的攻击场景，其要求对抗样本使得分类器分类为指定标签 t，如公式（2）所示。

$$\min \| \delta \| \ s.t. \ C(x+\delta) = t \tag{2}$$

其中扰动大小的度量公式常使用 P 范式，如公式（3）所示。本论文中使用 L_2 距离度量扰动的大小，即 2 范式、欧氏距离，其中 N 为像素点的个数。为了能够在连续的空间上搜索对抗样本，我们会将像素值从[0,255]压缩到[0,1]。同时,我们给出攻击成功的判定，一般为平均像素差值小于 0.05 且使分类器分类错误，对应到大小为(299,299,3)的图片上，L_2 距离约为 25。

$$\| x \|_p = \left(\sum_{i=1}^{N} | x_i |^p \right)^{\frac{1}{p}} \tag{3}$$

白盒对抗攻击最常用的就算是基于梯度的对抗样本生成方法，通过计算损失函数对输入图片的梯度，再通过梯度下降的方式降低损失值以达成目的。其中损失函数设计原理为损失值越小，对抗样本的效果越好，例如扰动的 L_2 距离加上目标分类置信度的负值。Szegedy 等人[6]首先设计了让神经网络做出误分类的最小扰动的方程，转而寻找最小的损失函数添加项，将对抗样本生成问题转化成了凸优化过程。Goodfellow 等人[7]使用 L_{inf} 距离设定扰动的上限，将原始分类的交叉熵作为目标函数，最大化目标函数以实现非目标攻击。Kurakin 等人[8]和 Makelov 等人[9]引入了迭代的思想，将扰动上限分成 n 份逐步完成攻击。Carlini 等人[11]将错误分类的约束条件统一至目标函数中，将对抗样本生成问题转化为优化问题，并设计了强有力的攻击方法。Papernot 等人[12]提出用梯度显著图的方式，每轮迭代仅选择效果最优的两个像素点进行修改，尽量降低扰动大小的同时完成了攻击。

1.3 黑盒对抗攻击

黑盒对抗攻击现在有两种常用的方法，TAEG 和 QAEG。

可转移性是对抗样本的一个很重要的性质。可转移性是指，针对某一个模型产生的对抗样本，可以也可以欺骗其他模型[29]。对抗样本存在可转移性主要原因是，同类型分类任务学到的分类区域可能大致相同，这导致对于一个模型有效的对抗样本可能对另一模型也有效。Papernot 等人[14]通过一定量的查询来构造与目标模型相似的替代模型，再使用白盒方法攻击这个替代模型以产生能够攻击目标黑盒模型的对抗样本。Dong 等人[10]将动量引入对抗样本生成过程，使得梯度更新的方向更加稳定，提高了攻击鲁棒性，同时提高了对抗样本的可转移性。Xie 等人[15]将输入变换引入到对抗样本生成过程，提高了对抗样本的可转移性。Tramer 等人[16]提出了联合多个模型进行生成以及防御对抗样本，大幅度提高了对抗样本的可转移性。Huang 等人[30]提出了针对模型的中间层而非 logit 层或输出层进行攻击，可以一定程度上提高对抗样本的可转移性。

尽管有众多学者对对抗样本的可转移性进行了研究，但是 TAEG 的黑盒成功率依然很低[15]，难以适应实际的需求。因此 QAEG 受到了越来越多的关注。QAEG 是通过查询目标模型，来估算模型对于输入的自然梯度，并辅以原始样本与目标图样本，逐步地进化对抗样本。在目标攻击场景下，QAEG 最常用基于攻击组的对抗攻击方法，即从目标图片出发，保持对抗样本的分类为目标分类，持续的降低对抗样本到原始图片的距离。Ilyas 等人[17]提出使用自然进化策略（Natural Evolutionary Strategies，NES[31]）来估算目标模型输出对抗样本分布期望的梯度，并通过进化整个分布来进化对抗样本。KUANG 等人[18]提出使用协方差矩阵自适应进化策略（Covariance Matrix Adaptation Evolutionary Strategies，CMA-ES[32]）来拟合并进化对抗样本的分布，并提升查询模型的效率，以降低完成攻击所需的查询次数。Brendel 等人[20]提出通过拒绝采样来拟合决策边界，不断的进化对抗样本，削弱了攻击算法对于模型自然梯度的依赖性，并实现了仅利用模型输出的 top1 标签完成攻击。Dong 等人[19]降低了搜索维度，并利用(1+1)-CMA-ES 来拟合决策边界分布，快速的进化对抗样本以实现攻击。

2 方法论

2.1 威胁模型

目前，攻击组的引入主要是为了解决黑盒模型下，目标攻击估算的自然梯度不精确的问题。因此本论文主要研究黑盒模型下的目标攻击的攻击组筛选方法。同时要求该黑盒模型可以被查询多次。

现有目标性对抗攻击主要有四种应用场景：

场景 1：原始分类固定，目标分类固定，但原始图片与目标图片均不固定。这是限定最弱的攻击场景，也是最常见的攻击场景，仅实现目标性攻击即可。实现人类识别结果和机器识别结果不同。例如，仅使目标分类系统出现错误。

场景 2：原始图像固定，目标分类固定，但目标图像不固定。这是另一种常用的攻击场景，给定原始图像，使得对抗样本与原始图像相差不大，但被分类器分类为目标分类。例如，自动驾驶欺骗。

场景 3：原始分类固定，目标图片固定，但原始图片不固定。这也是一种常见的攻击场景，例如，绕过人脸身份验证系统。

场景 4：原始图像固定，目标图像固定。这是最严格的攻击场景，但并不常见。

在前三种应用场景中，都存在很多符合要求的攻击组。本论文主要研究如何在前三种场景中，筛选出较为容易的攻击组，从而提高现有黑盒攻击的效率，以及攻击的稳定性。

2.2 基于攻击距离的对抗样本攻击组筛选方法

现有的 QAEG 虽然普遍使用了攻击组来指导攻击，但未注意到由于攻击组难度差异导致的攻击效果不稳定问题。因此目前也没有估算攻击组攻击难度的方法。为了解决以上问题，我们首先在 2.2.1 节提出了基于决策边界长度的攻击距离度量方法，然后在 2.2.2 节讨论了如何利用估算的攻击距离筛选对抗样本攻击组，最后通过预先筛选现有对抗样本生成方法的输入来大幅提升其攻击效率。基于攻击距离的对抗样本攻击组筛选方法的框架图如图 1 所示。

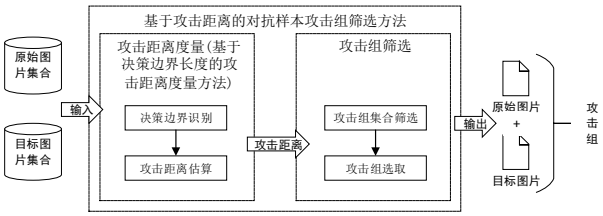

图 1 基于攻击距离的对抗样本攻击组筛选方法框架图

Fig.1 Framework of attack pairs filtering method based on attack distance

2.2.1 攻击距离度量

我们首先介绍对抗样本的迭代生成过程，再结合攻击速度图剖析攻击过程中的主要耗时部分，最

后总结并提出攻击距离度量方法以度量攻击组的攻击难度。

以 NES Attack 为例，其对抗样本生成的流程图如图 2 所示，攻击速度如图 3(a)所示。图 2 中 *s_img* 为原始图片，*t_img* 为目标图片，*adv_i* 是对抗样本的第 i 轮迭代结果。迭代的主要内容就是通过查询目标模型，并估算目标函数对输入的梯度，并通过梯度下降的方式更新对抗样本。同时，在攻击速度图中可以明显看到，对抗样本的 L_2 距离首先会快速的下降，然后下降曲线逐渐平缓。CMA 的攻击速度也是类似的，如图 3(b)所示。这其中的原因是，基于查询的攻击都会先快速找到决策边界，这个过程一般很快，然后再通过迭代稳步的更新对抗样本，这个过程一般很慢。而重要的是，第二个过程可以用线性拟合，以提前估算对抗样本生成所需要的查询次数。即查询次数和扰动的大小成反比，如果能够提前估算第二个阶段完成的攻击距离，就可以提前估算完成攻击所需的查询次数，从而估算攻击组的攻击难度。

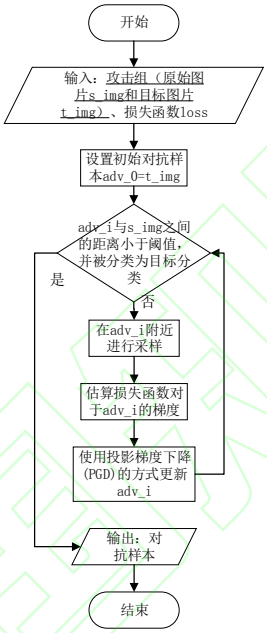

图 2　NES Attack 对抗样本生成流程图
Fig.2　Flow chart of NES Attack adversarial examples generation

图 3 展示了 NES Attack[17] 与 CMA Attack[18]在 ImageNet 上攻击 InceptionV3 的攻击速度图。两个方法针对同一攻击组的展开攻击，并展现出相似的攻击速度曲线。

基于查询的攻击过程有上述特性主要是由于对抗样本的特性。对抗样本是一种既要被分类为错误分类，又要与原始图片相差不大的样本。这就要求对抗样本一般都处于目标分类与原始分类的分类边界附近。因此对抗样本迭代生成的过程就是对抗样本游走于决策边界，并一步步向原始图片靠近的过程，包括基于梯度的白盒攻击与基于查询的黑盒攻击。估算一个攻击组完成攻击所需要的查询次数，主要需要解决两个问题：（1）如何确定对抗样本已经抵达决策边界附近。（2）如何估算攻击距离。

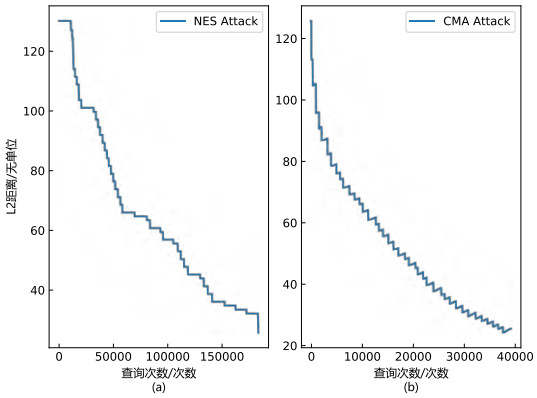

图 3 攻击速度图
Fig.3 Attack speed chart

为了解决第一个问题，我们设计了基于采样的决策边界识别方法，其效果示意图如图 4 所示。我们将原始图片与目标图片设为两个端点，并在两端点所在连线上进行二分查找，以快速定位决策边界位置。查找规则如下：如果大部分采样均为目标分类，则将中间图片向原始图片靠近；如果大部分采样均为原始分类，则将中间图片向目标图片靠近；直到在中间图片附近的采样有 40%~60%的概率为目标分类。图 4（a）展示了中间图片附近采样分类为目标分类的概率小于 40%的情况，中间图片将向目标图片靠近，即图 4（b）展现的情况。

为了解决第二个问题，我们设计了基于毕达哥拉斯定理的距离估算方法，其效果示意图如图 5 所示。设中间图片到原始图片的距离为 L_m，最终对抗样本到原始图片的距离为 L_d，从中间图片进化到对抗样本需要完成的攻击距离为 L_a。由于对抗样本进化的过程实际是在决策边界附近游走的过程，因此原始图片到最终对抗样本的距离可以看做是原始图片到决策边界的距离。又因为点到面垂线距离最短，因此最先找到符合 L_d 距离要求的对抗样本应该处于原始图片到决策边界的垂线附近。同时，对抗样本游走的攻击路径在决策边界附近，因此可以看做 L_d 近似垂直于 L_a。因此由毕达哥拉斯定理得，$L_a^2 = L_m^2 - L_d^2$。整个基于决策边界长度的攻击距离度量方法如算法 1 所示。

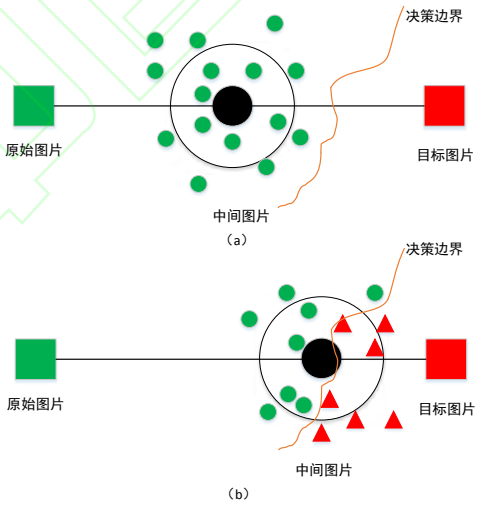

图 4 基于采样的决策边界识别方法示意图
Fig.4 Schematic diagram of decision boundary recognition method based on sampling

图 4、图 5 中，方块表示攻击组图片，黑色圆点表示中间图片，小圆点和小三角表示中间图片附近的采样；方块、圆点、三角的颜色表示其分类，绿色表示原始分类，红色表示目标分类；橙色的线表示决策边界。

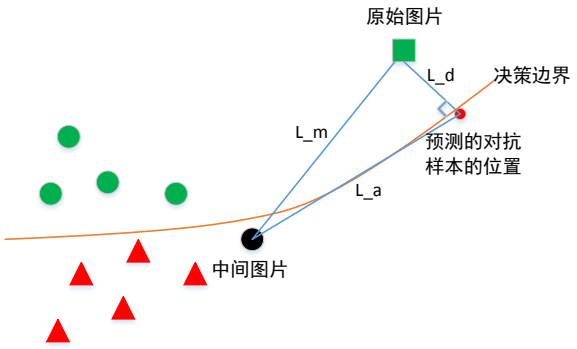

图 5　攻击距离估算示意图
Fig.5　Schematic diagram of attack distance estimation

算法 1　基于决策边界长度的攻击距离度量方法

输入：原始图片 *s_img*、目标图片 *t_img*、目标模型 C()、期望的对抗样本的扰动的大小 L_d
输出：攻击距离 L_a

1: *mid* = *t_img*, *low* = *s_img*, *high*=*t_img*// 图片的像素点范围为[0,255]
2: do :
3:　　X ~N (0,1) // N(0,1)为高维独立正态分布(以 ImageNet 上的 InceptionV3 为例，维数为 $299 \times 299 \times 3$)；X 为采样的集合，例如集合大小为 15
4:　　在 *mid* 附近生成采样 $X' = mid + X \times 2.55$
5:　　利用 C()计算 X'中对抗样本的比例 P
6:　　IF P >= 60%:
7:　　*mid* = (*mid* + *low*) / 2
8:　　*high* = *mid*
9:　　ELSE IF P <= 40%:
10:　　*mid* = (*mid* + *high*) / 2
11:　　*low* = *mid*
12:　　END IF
13:WHILE P>=60% || P <= 40%
14:$L_m = \|mid - s\_img\|_2$ // 此处 L_m 为中间图片到原始图片的 L2 距离，即欧式距离。
15: $L_a = \sqrt{L_m^2 - L_d^2}$
16:RETURN L_a

2.2.2 攻击组筛选

估算出攻击组之间的攻击距离之后我们就可以在现有攻击方法之前做一个预处理阶段，进行攻击组的筛选，以提高现有算法的攻击效果。

需要注意的是，攻击距离的估算结果并非十分精确，其能够大体的反映完成攻击所需要的查询次数，但也会存在一些偏差。因此不能直接利用估算的攻击距离选出最容易攻击的攻击组，但可以利用估算的攻击距离筛选出较容易攻击的攻击组。具体来说，我们从攻击组集合 S 中筛选出较容易攻击的攻击组集合 E，之后再从 E 中随机选取一组作为建议展开攻击的攻击组 *Pair*。三者之间的关系如公式（4）所示。

$$Pair \in E, E \subset S \tag{4}$$

其中，我们将集合 E 与集合 S 的比例定为超参数，即过滤因子 R。关于 R 值对于算法的影响，将在第 4 章进行详细的讨论。筛选时会首先会利用算法 1 估算 S 中各个攻击组的攻击距离，并将攻击组按照攻击距离升序排序。之后筛选前 $\eta = \rho * R$ 个攻击组来组成 E，其中 ρ 为 S 中攻击组的个数，η 是 E 中攻击组的个数。最后在 E 中随机选取一组作为 *Pair*。基于攻击距离的对抗样本攻击组筛选方法如算法 2 所示。

算法 2　基于攻击距离的对抗样本攻击组筛选方法

输入：原始图片集合 *s_imgs*、目标图片集合 *t_imgs*、过滤因子 R

输出：合适的攻击组 *Pair*

1: 生成攻击组集合

$$S = \{(s\_img, t\_img) \mid s\_img \in s\_imgs, t\_img \in t\_imgs\}$$

2: ρ = len(***S***)

3: 利用算法 1 获取 *S* 中每一组的攻击距离

4: 将攻击组按照攻击距离降序排序

5: $\eta = \rho * R$

6: 选取前 η 个攻击组组成 ***E***

7: *index* = rand(0, η) // 在 ***E*** 中随机选取一个攻击组

8: *Pair* = *E*[index]

9: RETURN *Pair*

3 实验与评估

本章节将讨论攻击组筛选方法对不同的攻击方法在不同应用场景下的有效性，同时讨论超参数 R 对于算法的影响。

本文主要对比的攻击方法包括 NES Attack[17]和 CMA Attack[18]，目标模型为 InceptionV3。我们将对比攻击组筛选前后实现攻击所需要查询次数的平均值、中位数、及其方差，从而验证攻击组筛选的有效性。

实验设定如下，选取 ImageNette 数据集进行验证与评估，一共 10 个分类，场景 1 时每种分类选取 10 张图片参与攻击组的构造，场景 2、3 时每种分类选取 100 张图片。场景 1 与场景 2、3 实验环境设置不同是因为场景 1 的组合情况过多，若每种分类选取图片均为 100 个，场景 1 的攻击组个数会达到 900,000 = 90×100×100，而场景 2、3 的攻击组个数仅为 9,000 = 90×100。因此降低场景 1 的每种分类选取的图片为 10 个，以平衡 3 种场景的攻击组个数均为 9,000 = 90×10×10。

具体来说，10 个分类一种有 90 种原始分类与目标分类不同的组合。场景 1 时，分类 A 中从 10 个图片选取 1 个作为原始图片，分类 2 中也从 10 个图片选取 1 个作为目标图片，一共有 100 种攻击组合，因此场景 1 一共有 9,000 种攻击组。场景 2 时，分类 A 中固定选取 1 张图片作为原始图片，分类 B 中从 100 个图片中选取 1 个作为目标图片，一共有 100 种攻击组合，因此场景 2 一共也有 9,000 种攻击组。场景 3 与场景 2 类似，分类 A 中从 100 个图片选取 1 个作为原始图片，分类 B 中固定选取 1 张图片作为目标图片，一共 100 中组合，因此场景 3 一共也有 9,000 种攻击组。

本文首先讨论不同 *R* 值，在场景 1 下对不同攻击算法的影响，如图 6 所示。随着 *R* 值的降低，两方法的平均查询次数与查询次数中位数均稳定的下降，验证了攻击组筛选对于提高查询效率的有效性。同时随着 *R* 值的降低，CMA Attack 的查询次数方差稳定下降，NES Attack 的查询次数方差波动下降，这验证了攻击组筛选对于稳定方差的有效性。实验结果显示，*R* 值越小，攻击效果越好，因此推荐将 *R* 值设为 0.1。本文之后的性能对比也是基于 *R* 值为 0.1 的基础上进行的。

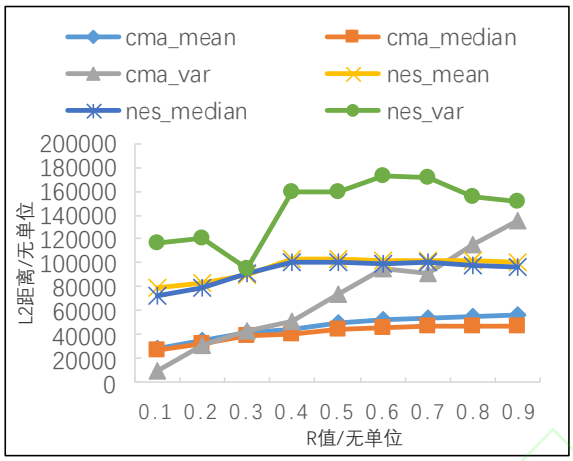

图 6　R 值影响图
Fig.6　Influence diagram of R's value

图 6 展现了不同的 *R* 值对于攻击组筛选方法的影响。其中 *cma_mean*、*cma_median*、*cma_var* 分别代表 CMA Attack 攻击成功所需查询次数的平均值、中位数、和方差；*nes_mean*、*nes_median*、*nes_var* 分别代表 NES Attack 攻击成功所需要查询次数的平均值、中位数、和方差；同时为了共用纵坐标轴，方差曲线显示的值为实际方差的 1/10000。

本文之后讨论攻击组筛选前后的攻击效果对比，如表 1 所示。表 1 展现了不同对抗样本生成方法在不同场景下实现攻击所需要的查询次数的平均值、中位值和方差。表 1 的第一列给出了所有待测的方法，包括 NES Attack[17]、CMA Attack[18]、经过攻击组筛选的 NES Attack（Filter NES）和经过攻击组筛选的 CMA Attack（Filter CMA）。表 1 的第 2 至 4 列给出了在场景 1（见 2.1）下，不同的对抗样本生成方法完成攻击所需要查询次数的平均值、中位值和方差。表 1 的第 5 至 7 列给出了在场景 2 下的情况，第 8-10 列给出了在场景 3 下的情况。

表 1 中倒数第二行给出了攻击组筛选在每一个分项上带来的平均提升效果。其计算公式如公式（4）所示。表 1 最后一行给出了攻击组筛选在不同场景、不同对抗样本生成方法下的平均提升效果。其中，总体提升效果表示攻击组筛选在不同场景下平均值、中位值和方差的综合提升，即将所有平均提升效果求平均：平均提升效率表示攻击组筛选在不同场景下的平均值、中位值的综合提升，即将不同场景下的平均值与中位值提升效果统一求平均；平均稳定方差表示攻击组筛选在不用场景下的平均稳定方差效果，即将不同场景下的方差提升效果求平均。

$$\text{平均提升效果}=\frac{\frac{NESAttack-FilteredNES}{NESAttack}+\frac{CMAAttack-FilteredCMA}{CMAAttack}}{2} \tag{5}$$

式中:NESAttack、FilteredNES、CMAAttack 和 FilteredCMA 分别为 NESAttack、Filtered NES、CMA Attack、Filtered CMA 完成攻击所需要的查询次数向量，分别对应表 1 中的第 3、4、5 和 6 行的数据。

场景 2 下我们的筛选效果不是很明显，这是可能是因为原始样本在筛选中占有更重要的地位。一旦固定原始样本的位置，那么原始样本附近的决策边界布局将确定，最后的优化难度也将基本确定。而越到最后的优化过程是越困难的，总体的优化难度更多取决于最后过程的优化难度。这也印证了一个普遍的实验现象，即对抗攻击的速度是先快后慢的。也就是说，如果在 A 类中选定了原始图片，与 B 类中的众多目标图片形成的攻击组具有相似的攻击距离。

但，无论是哪种场景下，攻击组筛选后的攻击效果都是优于筛选前的。筛选后攻击效率提高了 24.99%，方差稳定了 76.23%，总体攻击效果提升了 42.07%，充分验证了基于决策边界长度的攻击距离度量方法的有效性，验证了攻击组筛选的有效性。

最后关于本文所提方法通用性的讨论。本论文所提出的算法 1 与算法 2 是以图像识别领域为例，

但均不局限于图像识别领域。这是因为，算法 1 的适用前提有两个，（1）分类器模型的输入样本可以测量样本间距离，以及（2）可以在样本附近进行高斯采样。如果这两个前提条件能够满足，那么就很容易将算法 1 进行复现。例如语音识别领域的语音向量、恶意代码识别的代码向量[33]等。而算法 2 是基于算法 1 的，没有任何额外的限制。因此本文提出的方式是适用于所有领域的分类器的。

同时，本文所提出的方法适用所有对抗样本生成方法的攻击组的选取。这是由于我们的算法 1 与算法 2 的设计不局限于任何对抗样本生成方法，并在对比实验中证实了我们的方法对多种对抗样本生成方法有效。

表 1 筛选前后查询次数对比表

Table 1 Comparison of query times before and after filtering

方法	场景 1			场景 2			场景 3		
	平均值	中位数	方差	平均值	中位数	方差	平均值	中位数	方差
NES Attack	1.01e5	9.64e4	1.51e9	9.79e4	9.63e4	1.00e9	9.76e4	8.98e4	9.52e8
Filtered NES	7.87e4	7.22e4	1.16e9	9.66e4	9.66e4	4.20e8	6.64e4	6.64e4	3.59e6
CMA Attack	5.81e4	4.74e4	1.32e9	4.75e4	4.33e4	2.99e8	5.92e4	5.14e4	6.22e8
Filtered CMA	2.88e4	2.74e4	9.36e7	4.26e4	4.26e4	1.14e7	3.05e4	3.05e4	7.79e7
平均提升效果	36.25%	33.65%	58.04%	5.82%	0.65%	77.09%	40.22%	33.36%	93.55%
总体提升效果		42.07%	平均提升效率		24.99%		平均稳定方差		76.23%

4 结 论

1) 针对攻击组间攻击难度不同导致攻击不稳定的问题，本文以图像识别领域为例，设计了基于决策边界长度的攻击距离度量方法，为攻击组的攻击难度提供了度量方法。

2) 并在此基础上设计了基于攻击距离的对抗样本攻击组筛选方法，筛去难以攻击的攻击组。实现通过预处理，提升现有算法的攻击效率、攻击稳定性。

3) 实验表明，筛选后的攻击组相比于筛选前，攻击效率提升了 24.99%，方差稳定了 76.23%，总体攻击效果提升了 42.07%。

4) 最后，本文建议所有利用攻击组的对抗样本生成方法，先进行攻击组的筛选，特别是对原始样本进行筛选，再展开攻击，以稳定攻击的效果。

同时，为了更加清晰筛选对于攻击组的意义，我们将会继续研究原始样本对于整个优化过程的意义。

参考文献（References）

[1] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]//Advances in neural information processing systems. 2012: 1097-1105.

[2] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 1-9.

[3] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.

[4] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[5] Huang G, Liu Z, Van Der Maaten L, et al. Densely connected convolutional networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 4700-4708.

[6] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks[J]. arXiv preprint arXiv:1312.6199, 2013.

[7] Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples[J]. arXiv preprint arXiv:1412.6572, 2014.

[8] Kurakin A, Goodfellow I, Bengio S. Adversarial machine learning at scale[J]. arXiv preprint arXiv:1611.01236, 2016.

[9] Madry A, Makelov A, Schmidt L, et al. Towards deep learning models resistant to adversarial attacks[J]. arXiv preprint arXiv: 1706.06083, 2017.

[10] Dong Y, Liao F, Pang T, et al. Boosting adversarial attacks with momentum[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 9185-9193.

[11] Carlini N, Wagner D. Towards evaluating the robustness of neural networks[C]//2017 ieee symposium on security and privacy (sp). IEEE, 2017: 39-57.

[12] Papernot N, McDaniel P, Jha S, et al. The limitations of deep learning in adversarial settings[C]//2016 IEEE European symposium on security and privacy (EuroS&P). IEEE, 2016: 372-387.

[13] Moosavi-Dezfooli S M, Fawzi A, Frossard P. Deepfool: a simple and accurate method to fool deep neural networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2574-2582.

[14] Papernot N, McDaniel P, Goodfellow I, et al. Practical black-box attacks against deep learning systems using adversarial examples[J]. arXiv preprint arXiv:1602.02697, 2016, 1(2): 3.

[15] Xie C, Zhang Z, Zhou Y, et al. Improving transferability of adversarial examples with input diversity[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 2730-2739.

[16] Tramèr F, Kurakin A, Papernot N, et al. Ensemble adversarial training: Attacks and defenses[J]. arXiv preprint arXiv:1705.07204, 2017.

[17] Ilyas A, Engstrom L, Athalye A, et al. Black-box adversarial attacks with limited queries and information[J]. arXiv preprint arXiv:1804.08598, 2018.
[18] Kuang X, Liu H, Wang Y, et al. A CMA-ES-Based Adversarial Attack on Black-Box Deep Neural Networks[J]. IEEE Access, 2019, 7: 172938-172947.
[19] Dong Y, Su H, Wu B, et al. Efficient decision-based black-box adversarial attacks on face recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 7714-7722.
[20]] Brendel W, Rauber J, Bethge M. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models[J]. arXiv preprint arXiv:1712.04248, 2017.
[21] LeCun Y, Bengio Y, Hinton G. Deep learning[J]. nature, 2015, 521(7553): 436-444.
[22] Helmstaedter M, Briggman K L, Turaga S C, et al. Connectomic reconstruction of the inner plexiform layer in the mouse retina[J]. Nature, 2013, 500(7461): 168-174.
[23] Xiong H Y, Alipanahi B, Lee L J, et al. The human splicing code reveals new insights into the genetic determinants of disease[J]. Science, 2015, 347(6218): 1254806.
[24] Hinton G, Deng L, Yu D, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups[J]. IEEE Signal processing magazine, 2012, 29(6): 82-97.
[25] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks[C]//Advances in neural information processing systems. 2014: 3104-3112.
[26] Mnih V, Kavukcuoglu K, Silver D, et al. Human-level control through deep reinforcement learning[J]. Nature, 2015, 518(7540): 529-533.
[27] Giusti A, Guzzi J, Cireşan D C, et al. A machine learning approach to visual perception of forest trails for mobile robots[J]. IEEE Robotics and Automation Letters, 2015, 1(2): 661-667.
[28] Middlehurst C. China unveils world's first facial recognition atm[EB/OL]. London, England: The Telegraph, 2015(2015-01-01)[2020-09-01]. https://www.telegraph.co.uk/news/worldnews/asia/china/11643314/China-unveils-worlds-first-facial-recognition-ATM.html
[29] 刘恒，吴德鑫，徐剑．基于生成式对抗网络的通用性对抗扰动生成方法[J]．信息网络安全，2020，20（5）：57-64．
LIU H, WU D, XU J. Generating Universal Adversarial Perturbations with Generative Adversarial Networks[J]. Netinfo Security, 2020, 20(5): 57-64(in Chinese).
[30] Huang Q, Katsman I, He H, et al. Enhancing adversarial example transferability with an intermediate level attack[C]//Proceedings of the IEEE International Conference on Computer Vision. 2019: 4733-4742.
[31] Wierstra D, Schaul T, Peters J, et al. Natural evolution strategies[C]//2008 IEEE Congress on Evolutionary Computation (IEEE World Congress on Computational Intelligence). IEEE, 2008: 3381-3387.
[32] Hansen N. The CMA evolution strategy: A tutorial[J]. arXiv preprint arXiv:1604.00772, 2016.
[33] 侯留洋，罗森林，潘丽敏，等．融合多特征的 Android 恶意软件检测方法[J]．信息网络安全，2020，20（1）：67-74．
HOU L, LUO S, PAN L, et. Multi-feature Android Malware Detection Method[J]. Netinfo Security, 2020, 20(1): 67-74(in Chinese).