

面向对抗样本攻击的移动目标防御

王滨^{1,2,3}, 陈靓¹, 钱亚冠¹, 郭艳凯¹, 邵琦琦¹, 王佳敏¹

(1. 浙江科技学院大数据学院, 浙江 杭州 310023;

2. 浙江大学电气工程学院, 浙江 杭州 310058;

3. 海康威视数字技术有限公司网络与信息安全实验室, 浙江 杭州 310058)

摘要: 深度神经网络已被成功应用于图像分类, 但研究表明深度神经网络容易受到对抗样本的攻击。提出一种移动目标防御方法, 通过 Bayes-Stackelberg 博弈策略动态切换模型, 使攻击者无法持续获得一致信息, 从而阻断其构建对抗样本。成员模型的差异性是提高移动目标防御效果的关键, 将成员模型之间的梯度一致性作为度量, 构建新的损失函数进行训练, 可有效提高成员模型之间的差异性。实验结果表明, 所提出的方法能够提高图像分类系统的移动目标防御性能, 显著降低对抗样本的攻击成功率。

关键词: 对抗样本; 移动目标防御; Bayes-Stackelberg 博弈

中图分类号: TP393

文献标识码: A

doi: 10.11959/j.issn.2096-109x.2021012

Moving target defense against adversarial attacks

WANG Bin^{1,2,3}, CHEN Liang¹, QIAN Yaguan¹, GUO Yankai¹, SHAO Qiqi¹, WANG Jiamin¹

1. College of Science, Zhejiang University of Science and Technology, Hangzhou 310023, China

2. College of Electrical Engineering, Zhejiang University, Hangzhou 310058, China

3. Network and Information Security Laboratory, Hangzhou Hikvision Digital Technology Co., LTD, Hangzhou 310058, China

Abstract: Deep neural network has been successfully applied to image classification, but recent research work shows that deep neural network is vulnerable to adversarial attacks. A moving target defense method was proposed by means of dynamic switching model with a Bayes-Stackelberg game strategy, which could prevent an attacker from continuously obtaining consistent information and thus blocked its construction of adversarial examples. To improve the defense effect of the proposed method, the gradient consistency among the member models was taken as a measure to construct a new loss function in training for improving the difference among the member models. Experimental results show that the proposed method can improve the moving target defense performance of the

收稿日期: 2020-09-18; **修回日期:** 2020-12-08

通信作者: 钱亚冠, qianyaguan@zust.edu.cn

基金项目: 国家重点研发计划 (2018YFB2100400); 国家电网公司总部科技项目 (5700-202019187A-0-0-00); 2019 年度杭州市领军型创新团队项目

Foundation Items: The National Key R&D Program of China (2018YFB2100400), Science and Technology Project of State Grid Corporation of China (5700-202019187A-0-0-00), Hangzhou City Leading Innovation Team Project in 2019

论文引用格式: 王滨, 陈靓, 钱亚冠, 等. 面向对抗样本攻击的移动目标防御[J]. 网络与信息安全学报, 2021, 7(1): 113-120.

WANG B, CHEN L, QIAN Y G, et al. Moving target defense against adversarial attacks[J]. Chinese Journal of Network and Information Security, 2021, 7(1): 113-120.

image classification system and significantly reduce the attack success rate against the adversarial examples.

Keywords: adversarial examples, moving target defense, Bayes-Stackelberg game

1 引言

随着人工智能的发展, 基于深度神经网络(DNN)的图像分类技术被成功应用到多种任务中, 如手写字体识别^[1]、自动监控的对象分类^[2]和自动驾驶^[3]等。最近的研究表明, DNN 存在严重的安全脆弱性, 即在输入图像上添加一些精心设计的、不被人眼察觉的扰动, 可欺骗 DNN 错误分类。这种添加了恶意扰动的样本被称为对抗样本^[4]。对抗样本的出现限制了 DNN 在安全敏感领域的广泛应用。为了更好地防御对抗样本, 研究人员提出对抗训练防御方法^[5], 将对抗样本注入训练数据, 从而增强 DNN 的鲁棒性。Madry 等^[6]将对抗训练建模为最小-最大化对抗博弈, 攻击者生成对抗样本最大化损失函数, 而防守者在对抗样本存在的情况下最小化损失函数, 从而提高模型的鲁棒性。

对抗训练方法需要生成对抗样本对分类模型进行数据增强。与对抗训练防御对抗样本不同, Sengupta 等^[7]借鉴移动目标防御^[8](MTD, moving target defense)的思想, 将多个 DNN 模型构建成基于 MTD 的图像分类系统, 在运行过程中主动切换成员模型, 以降低攻击者的攻击成功率, 从而增强图像分类系统的鲁棒性。MTD 主要是利用特定攻击在不同的成员模型上的攻击差异性, 通过随机切换成员模型以增加攻击难度。但该工作的不足是: 系统成员模型只使用简单的训练数据划分, 选择不同的网络结构和权重参数等方法, 本文的实验表明上述方法并不能有效地提高系统模型之间的差异性, 因而实际的防御效果不佳。

本文围绕如何提高模型之间的差异性这一关键问题, 提出了差异性训练方法, 利用模型梯度之间的相关性来衡量模型差异性, 以进一步改进 MTD 的性能。为此, 构造新的损失函数, 在成员模型的平均交叉熵损失的基础上, 增加模型差异性对数项, 在最小化损失函数的同时, 增大成员模型间的差异性。在此基础上, 再利用 Bayes-Stackelberg 博弈, 获得模型切换的最佳策

略, 从而有效地阻止对抗样本攻击。本文提出的改进移动目标防御被称为 DT-MTD (differential training MTD), 并在 ILSVRC2012^[9]的竞赛数据集上验证了差异性训练的有效性。实验结果表明, 相比于未使用差异性训练的图像分类系统, 本文提出的 DT-MTD 可更进一步提高图像分类系统的鲁棒性。

2 预备知识

2.1 深度神经网络

深度学习模型可以表示成函数 $\mathcal{F}(x; \theta)$: $\mathcal{R}^d \rightarrow \mathcal{R}^L$, 其中 $x \in \mathcal{R}^d$ 是输入样本, θ 表示 DNN 的参数, L 表示 DNN 预测类别数量。本文采用带有 softmax 输出层 DNN, 其中 softmax 函数定义为

$$\text{softmax}(z_i) = \frac{\exp(z_i)}{\sum_{m=1}^L \exp(z_m)} \quad (1)$$

因此, DNN 可以表示为 $\mathcal{F}(x) = \text{softmax}(z)$, 其中 z 表示 DNN 最后一层隐藏层的输出向量。给定一个输入样本 x , DNN 的预测标签可以表示为: $y' = \arg\max_i \mathcal{F}(x)_i$, 其中概率值 $\mathcal{F}(x)_i$ 被称为第 i 类的置信度得分。训练 DNN 的目标就是让预测 y' 与真实标签 y 之间的差值越来越小, 用 $J(x, y; \theta)$ 来表示, 又称损失函数。本文采用交叉熵损失函数:

$$J(x, y; \theta) = -1_y \cdot \log(\text{Softmax}(z(x, y; \theta))) \quad (2)$$

其中, -1_y 是真实标签 y 的 one-hot 编码, 向量的对数定义为每个元素的对数。

2.2 对抗样本和威胁模型

对抗样本是在干净样本 x 中添加人眼无法察觉的扰动 r , 使 DNN 错误分类。对抗样本表示为 $x_{\text{adv}} = x + r$, 约束条件为 $\|r\|_p \leq \epsilon$, $\arg\max_i \mathcal{F}(x)_i \neq y$ 。本文使用 l_∞ 范数来表示扰动 r 的大小, 即 $\|r\|_\infty \leq \epsilon$ 。

在对抗环境中, 根据攻击者获取目标 DNN 内部信息的能力, 分为黑盒攻击和白盒攻击。①黑盒攻击: 攻击者对于目标 DNN 的信息一无所

知,但可以在与 DNN 的交互过程中获得模型的相关信息,如预测标签或置信度向量。②白盒攻击:攻击者完全了解目标 DNN 的信息,包括训练数据、模型参数、拓扑结构、训练优化方法等。本文假设的威胁模型是攻击者已知图像分类系统中的成员模型的相关信息,即攻击者可以构建与图像分类系统一样的代理模型 $\{\mathcal{F}_a(\theta^{(k)})\}_{k=1}^N$ 。为了增加攻击者选择代理模型的难度,MTD 可同时增加黑盒和白盒攻击的难度。

2.3 基于梯度生成对抗样本

当已知 DNN 的内部信息,白盒攻击可以在约束 $\|x_{\text{adv}} - x\|_p \leq \varepsilon$ 下,通过优化函数 $\arg\max_{x_{\text{adv}}} J(x_{\text{adv}}, y; \theta)$ 来构造对抗样本。本节介绍基于梯度优化生成对抗样本的攻击方法。

(1) FGSM

GoodFellow 等^[5]提出了第一种基于模型梯度信息生成对抗样本的方法,通过最大化损失函数 $J(x, y, \theta)$ 获取对抗样本 x_{adv} 。为此,在损失函数对于 x 的梯度变化最大的方向上寻找扰动 r ,即 $x_{\text{adv}} = x + r \cdot \text{sign}(\nabla_x J(x_{\text{adv}}, y; \theta))$,其中 $\text{sign}(\cdot)$ 表示符号函数, $\nabla_x J(x_{\text{adv}}, y; \theta)$ 是损失函数对于输入 x 的梯度, $\|r\|_\infty \leq \varepsilon$ 。

(2) PGD

Madry 等^[6]将 FGSM 扩展为迭代的方式寻找扰动,即 $x_{\text{adv}}^{t+1} = \text{Proj}(x_{\text{adv}}^t + \alpha \cdot \text{sign}(\nabla_x J(x_{\text{adv}}^t, y; \theta)))$,其中, t 是当前迭代次数,迭代步长 $\alpha = \frac{\varepsilon}{T}$, T 是整个迭代的总次数, $\text{Proj}(\cdot)$ 将每次更新的对抗样本投影到 x 的 ε 邻域内。

(3) MI-FGSM

Dong 等^[10]将 PGD 的迭代部分用动量迭代代替,以稳定梯度方向以免进入局部最大值。基于动量的梯度下降的多步迭代方法表示为 $x_{\text{adv}}^{t+1} = \text{clip}_\varepsilon(x_{\text{adv}}^t + u \text{sign}(g_t))$,其中, $g_{t+1} = u g_t + \frac{\nabla_x J(x_{\text{adv}}^t, y; \theta)}{\|\nabla_x J(x_{\text{adv}}^t, y; \theta)\|_2}$, u 是动量项衰减因子。

(4) M-DI²-FGSM

Xie 等^[11]为了提高多步迭代方法的黑盒攻击成功率,提出了在每一步迭代完成后对样本进行

输入变换,即 $x_{\text{adv}}^{t+1} = \text{clip}_\varepsilon(x_{\text{adv}}^t + \alpha \text{sign}(\nabla_x J(T(x_{\text{adv}}^t, p), y; \theta)))$,其中 p 表示变换概率,随机变换函数

$$T(x, p) = \begin{cases} T(x), & p \\ x, & 1-p \end{cases}.$$

3 防御方法

本节介绍本文提出的差异性训练方法,在此基础上利用 Bayes-Stackelberg 博弈切换 DNN,混淆攻击者获得的信息,使其难以选择合适的代理模型,从而提高图像分类系统的鲁棒性。

3.1 差异性训练

移动目标防御成功的关键是特定攻击在图像分类系统的成员模型上具有差异性,这种攻击差异性的定义为

$$DI = \min_{\mathcal{F}_a} \frac{\max_{\mathcal{F}_d} E(\mathcal{F}_a, \mathcal{F}_d) - \min_{\mathcal{F}_d} E(\mathcal{F}_a, \mathcal{F}_d) + 1}{\max_{\mathcal{F}_d} E(\mathcal{F}_a, \mathcal{F}_d) + 1} \quad (3)$$

其中, $\mathcal{F}_a \in M^A$ 表示攻击者选择的一个代理模型, M^A 表示一组代理模型, $\mathcal{F}_d \in N$ 表示防御方通过博弈选择的系统成员模型, N 表示系统成员模型集合, $E(\mathcal{F}_a, \mathcal{F}_d)$ 表示 $\mathcal{F}_a$ 上生成的对抗样本攻击 $\mathcal{F}_d$ 的成功率。 DI 值越大,意味着 DNN 移动目标防御的效果越好。

本文通过增大移动目标防御系统中成员模型之间的梯度差异性,从而增大 DI 值。假设有成员模型 $\mathcal{F}_d^1$ 和 $\mathcal{F}_d^2$ 以及攻击者选择的代理模型 $\mathcal{F}_a$,用 $\nabla_x J_1$ 和 $\nabla_x J_2$ 分别表示 $\mathcal{F}_d^1$ 和 $\mathcal{F}_d^2$ 的代价函数对样本 x 的梯度, $\nabla_x J_a$ 表示 $\mathcal{F}_a$ 的代价函数对样本 x 的梯度, $\mathcal{F}_a$ 生成对抗样本 x_{adv} 。如果 $\nabla_x J_1$ 和 $\nabla_x J_2$ 的方向一致,这意味着能使 $\mathcal{F}_d^1$ 误分类的 x_{adv} 也能使 $\mathcal{F}_d^2$ 误分类,那么 $\mathcal{F}_d^1$ 和 $\mathcal{F}_d^2$ 之间的差异性较弱。当代理模型 $\mathcal{F}_a$ 的代价梯度 $\nabla_x J_a$ 一定时, $\nabla_x J_1$ 和 $\nabla_x J_2$ 梯度越对齐, $\mathcal{F}_d^1$ 、 $\mathcal{F}_d^2$ 和 $\mathcal{F}_a$ 的对抗子空间重叠区域越大。本文使用余弦相似度来表示 $\nabla_x J_1$ 和 $\nabla_x J_2$ 的对齐程度,如式(4)所示。

$$CS(\nabla_x J_1, \nabla_x J_2) = \frac{\langle \nabla_x J_1, \nabla_x J_2 \rangle}{\|\nabla_x J_1\| \cdot \|\nabla_x J_2\|} \quad (4)$$

其中, $CS(\cdot) \in [-1, 1]$ 。若 $CS(\nabla_x J_1, \nabla_x J_2)$, 则 $\nabla_x J_1$

和 $\nabla_x J_2$ 的梯度方向相反, 意味着对抗样本 x_{adv} 如果能攻击成功 $\mathcal{F}_d^1$, 就不能攻击成功 $\mathcal{F}_d^2$, 这会导致式(3)中分子 $\max_{\mathcal{F}_d} E(\mathcal{F}_a, \mathcal{F}_d) - \min_{\mathcal{F}_d} E(\mathcal{F}_a, \mathcal{F}_d)$ 项增大, 从而增大 $\mathcal{DI}$ 。对于 K 个成员模型的图像分类系统, 可以取成员模型之间最大的 CS 值, 如式(5)所示。

$$M_{\text{CS}} = \max_{a, b \in \{1, \dots, K\}, a \neq b} \text{CS}(\nabla_x J_a(\theta_a), \nabla_x J_b(\theta_b)) \quad (5)$$

为了增大 $\mathcal{DI}$, 最大化 M_{CS} , 但由于式(5)是非光滑函数, 无法使用梯度下降等一阶优化方法, 因此, 对式(5)使用 LogSumExp 函数进行光滑逼近。

$$\psi_{\text{CS}} = \log\left(\sum_{1 \leq a < b \leq K} \exp(\text{CS}(\nabla_x J_a(\theta_a), \nabla_x J_b(\theta_b)))\right) \quad (6)$$

如果 K 个模型的 ψ_{CS} 较小, 那么模型之间就具有较大的差异性, 从而增大 $\mathcal{DI}$ 。为此, 将 ψ_{CS} 作为正则化项加入 K 个模型集成训练的交叉熵损失函数中, 如式(7)所示。

$$\text{Loss} = \frac{1}{K} \sum_{k=1}^K J(x, y; \theta_k) + \lambda \cdot \psi_{\text{CS}} \quad (7)$$

其中 λ 作为超参数控制在训练过程中 ψ_{CS} 的比重。由于模型训练的目的是使目标函数 Loss 最小, 因此当增大 λ 的值时, 必然会减小 ψ_{CS} , 即增大模型之间的差异性; 反之, 减小 λ 的值时, 必然会增大 ψ_{CS} , 即降低了模型之间的差异性。

3.2 基于 Bayes-Stackelberg 博弈的模型切换

Stackelberg 博弈^[12]是一种非合作的、有先后次序的决策博弈, 参与者包括先采取行动的领导者 (leader) 以及随后启动的跟随者 (follower)。领导者和跟随者各自具有一个行动集合, 参与者从集合中选择一个行动形成策略。领导者首先发布一个混合策略, 追随者在领导者的策略下优化自身收益, 回应一个纯策略。Bayes-Stackelberg 博弈是将 Stackelberg 博弈扩展为多个类型的追随者, 适用于信息安全中的博弈建模。把它应用到 DT-MTD 中, 求解得到贝叶斯-纳什均衡, 并将其作为最佳的模型切换策略。本文的领导者为图像分类系统的防御者, 追随者为用户 (包括正常用户和攻击者), 防御者和追随者的目标都是最大化自己的收益。

本文将 Bayes-Stackelberg 博弈模型定义为四

元组 $G_{\text{BSG}} = (N, T, M, U)$ 。其中 $N = (n_1, n_2, \dots, n_k)$ 是防御者的行动集合, 在本文中表示 DT-MTD 成员模型 $N = \{\mathcal{F}(\theta_k)\}_{k=1}^K$ 的选择。 $T = (t_1, t_2, \dots, t_i)$ 是用户类型的集合, $T \neq \emptyset$, 本文包含两种用户类型, 合法用户 $\mathcal{L}$ 和攻击者 $\mathcal{A}$ 。用 α 表示攻击者 $\mathcal{A}$ 出现的概率, $1-\alpha$ 表示合法用户 $\mathcal{L}$ 出现的概率。 $M = (M^1, M^2, \dots, M^i)$, 对任意的 $M^i \neq \emptyset$, $M^i = (m_1, m_2, \dots, m_h)$, 表示每种类型的用户至少有一个以上的策略。合法用户 $\mathcal{L}$ 只有一种策略, 即使用干净图像请求分类服务。而对于攻击者 $\mathcal{A}$ 来说, 其行动集合 $M^{\mathcal{A}} = \{\mathcal{F}_d(\theta_i)\}_{i=1}^h$ 是对一组代理模型的选择。 $U=(R, C)$ 是收益函数集合, R 和 C 分别表示防御者和用户的收益。对于给定的用户类型 t 、纯策略 q^t 和防御者混合策略 s , 双方的期望收益如式(8)、式(9)所示。

$$R(s, q^t, t) = \sum_{n \in N} \sum_{m \in M^t} R_{nm}^t s_n q_m^t \quad (8)$$

$$C(s, q^t, t) = \sum_{n \in N} \sum_{m \in M^t} C_{nm}^t s_n q_m^t \quad (9)$$

其中, R_{nm}^t 表示防御者在选择策略 n , 用户 t 选择策略 m 时防御者的收益, C_{nm}^t 表示用户 t 的收益。本文将防御者和合法用户的收益定义为干净图像的分类准确率, 对于攻击者, 其收益定义为攻击成功率。

在该博弈中, 防御者策略 s_n 需要满足约束 $\sum_{n \in N} s_n = 1$, $s_n \in [0, 1]$, 表示防御者根据策略 s 选择成员模型。用户策略需要满足 $\sum_{m \in M^t} q_m^t = 1$, $q_m^t \in \{0, 1\}$,

对于防御者确定的混合策略调度系统成员模型, 每种类型的用户面对的是一个线性收益问题, 存在一个最优策略, 故每种用户都会选择一种纯策略回应。用户的收益约束如式(10)所示。

$$0 \leq (v^t - \sum_{n \in N} C_{nm} s_n) \leq (1 - q_m^t) D, \quad v^t \in R \quad (10)$$

其中, D 是一个大的正数。当 $q_m^t = 0$ 时, 式(10)右边的不等号是不成立的; 当 $q_m^t = 1$ 时左边的不等式使用户的收益最大化, 当用户选择的策略最大化自身收益时, 两边取等号。综上可知, 通过博弈, 将系统成员模型的调度策略问题, 转化为一个混合整数二次规划问题 (MIQP), 目标函数的最优解 s 即为最优混合调度策略, 如式(11)所示。

$$\begin{aligned}
& \max_{s,q,v} \sum_{n \in N} (\alpha \cdot \sum_{m \in M^A} R_{n,m}^t S_n q_m^A + (1-\alpha) \cdot R_{n,m}^t S_n q_n^C) \\
& \text{约束条件为 } \sum_{n \in N} s_n = 1, 0 \leq s_n \leq 1 \\
& \sum_{m \in M^t} q_m^t = 1, q_m^t \in \{0,1\} \\
& 0 \leq v^t - \sum_{n \in N} C_{n,m}^t S_n \leq (1 - q_m^t) D \\
& v^t \in R \\
& \forall t \in \{\mathcal{L}, \mathcal{A}\}
\end{aligned} \quad (11)$$

其中, α 表示用户类型是攻击者 $\mathcal{A}$ 的概率。本文利用 DOBSS 算法^[13]求解式(11), 此时图像分类系统成员模型的最优切换策略是 s, q 是用户回应的最优策略, v 是用户的收益。

4 实验评估

4.1 实验设置

本文实验使用 ILSVRC2012 的竞赛数据集, 是 ImageNet^[14]的一个子集。将该数据集中的 1.2×10^6 张图像作为训练集, 50 000 张图像作为验证集, 以及 150 000 张图像作为测试集, 包含 1 000 个类别标注, 本文使用训练集训练 DT-MTD 的成员模型, 在测试集上进行验证, 攻击者使用测试集生成对抗样本。实验使用带 4 个 NVIDIA GeForce RTX 2080 Ti GPU 的服务器, 该服务器的操作系统是 Ubuntu16.04.6 LTS, 使用 Python3.6、Ptorch1.2、PuLP2.1 实现差异性训练和 DOBSS 算法的求解。

实验中的 DT-MTD 成员模型结构是 { GoogLeNet^[15], VGG16^[16], ResNet-50^[17] }, 利用算法 1 进行差异性训练。表 1 是防御者与合法用户在 ILSVRC2012 数据集的测试集上的收益, 如表 1 中的 (91.3, 91.3) 表示当防御者使用成员模型 GoogLeNet 对干净图像进行分类时, 防御者的收益和用户收益都是分类准确率 91.3%。

算法 1 差异性训练算法

输入 待训练模型 $\{\mathcal{F}(\theta_k)\}_{k=1}^K$; 训练数据 $D = \{(x_i, y_i)\}_{i=1}^N$; 学习率 $\{\eta_k\}_{k=1}^K$; 训练轮数 N

1) 初始化模型权重 $\{\theta_k\}_{k=1}^K$

2) for $n = 1, \dots, N$ do

3) $\text{Loss}_n \leftarrow \frac{1}{|D|} \sum_{(x,y) \in D} \left(\frac{1}{K} \sum_{k=1}^K J(x, y; \theta_k) + \lambda \psi_{CS} \right)$

//计算损失

4) for $k = 1, \dots, K$ do

5) $\theta_k^{m+1} \leftarrow \theta_k^m - \eta_k \cdot \nabla_{\theta_k^m} \text{Loss}_n$

//更新每个成员模型的权重

6) end for

7) end for

8) return $\{\mathcal{F}(\theta_k)\}_{k=1}^K$

表 1 防御者和合法用户的博弈收益
Table 1 Game payoffs of defenders and legitimate users

GoogLeNet ^[15]	VGG16 ^[16]	ResNet-50 ^[17]
(91.3, 91.3)	(85.5, 85.5)	(92.5, 92.5)

此外, 实验中生成对抗样本的扰动约束 $\varepsilon=10$; PGD 从均匀分布 $U(-\varepsilon, +\varepsilon)$ 初始化扰动, 迭代 20 次, 扰动步长为 2; 对于 MI-FGSM 和 M-DI²-FGSM, 迭代 10 次, 扰动步长为 $\frac{\varepsilon}{10}$, 动量项衰减因子 $\mu=1$; M-DI²-FGSM 中的变换操作 $T(x, \rho)$, 概率 $\rho=0.5$, 变换操作的具体方式是在模型输入前, 将输入 x 随机调整为 $\text{rand} \times \text{rand} \times 3$ 图像, 其中 $\text{rand} \in [299, 300]$ 。

根据假设的威胁模型, 攻击者知道 DT-MTD 的成员模型信息, 可以构建出代理模型集合 { GoogLeNet, VGG16, ResNet-50 }。表 2~表 5 是防御者和不同的攻击者 (FGSM、PGD、MI-FGSM 和 M-DI²-FGSM) 之间的收益矩阵。以攻击者使用 PGD 生成对抗样本为例, 表 3 是防御者与 PGD 攻击者的博弈收益。表 3 中数据 (37.5, 62.5) 表示攻击者在代理模型 GoogLeNet 上使用 PGD 生成对抗样本, 攻击成员分类模型 VGG16 时, 防御者的收益为对抗样本的分类准确率 37.5%, 攻击者的收益是对抗样本攻击成功率 62.5%。从表 3 还可以发现基于特定代理模型生成的对抗样本, 在不同成员模型上表现出攻击差异性, 如表 3 中的第 3 列, 代理模型 VGG16 基于 PGD 生成的对抗样本, 对 GoogLeNet 的攻击成功率为 41.2%, 而对 ResNet-50 的攻击成功率为 52.0%。这说明使用移动目标防御技术, 切换成员模型可以降低攻击者的攻击成功率。

表 2 防御者和 FGSM 攻击者的博弈收益
Table 2 Game payoffs of defenders and FGSM attackers

防御者	FGSM 攻击者		
	GoogLeNet	VGG16	ResNet-50
GoogLeNet	(21.1, 79.9)	(62.5, 37.5)	(55.1, 44.9)
VGG16	(40.5, 59.5)	(26.7, 73.3)	(33.4, 66.6)
ResNet-50	(43.6, 53.4)	51.1, 48.9	(16.3, 83.7)

表 3 防御者和 PGD 攻击者的博弈收益
Table 3 Game payoffs of defenders and PGD attackers

防御者	PGD 攻击者		
	GoogLeNet	VGG16	ResNet-50
GoogLeNet	(18.1, 81.9)	(38.8, 41.2)	(52.5, 47.5)
VGG16	(37.5, 62.5)	(19.7, 80.3)	(30.6, 69.4)
ResNet-50	(49.5, 50.5)	(48.0, 52.0)	(14.0, 86.0)

表 4 防御者和 MI-FGSM 攻击者的博弈收益
Table 4 Game payoffs of defenders and MI-FGSM attackers

防御者	MI-FGSM 攻击者		
	GoogLeNet	VGG16	ResNet-50
GoogLeNet	(14.1, 85.9)	(60.8, 39.2)	(54.8, 45.2)
VGG16	(45.7, 54.3)	(21.3, 78.7)	(36.6, 63.4)
ResNet-50	(48.5, 51.5)	(52.2, 47.8)	(15.4, 84.6)

表 5 防御者和 M-DI²-FGSM 攻击者的博弈收益
Table 5 Game payoffs of defenders and M-DI²-FGSM attackers

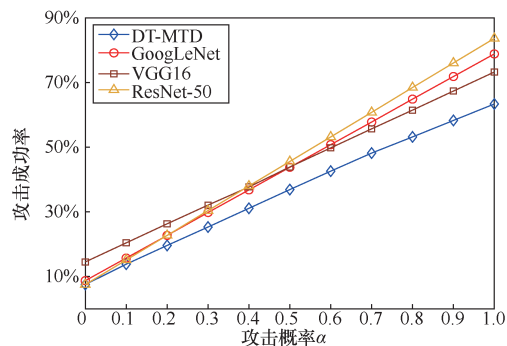
防御者	M-DI ² -FGSM 攻击者		
	GoogLeNet	VGG16	ResNet-50
GoogLeNet	(18.8, 81.2)	(61.1, 38.9)	(54.0, 46.0)
VGG16	(38.6, 61.4)	(24.2, 75.8)	(31.8, 68.2)
ResNet-50	(47.9, 52.1)	(50.1, 49.9)	(14.3, 85.7)

4.2 DT-MTD 的鲁棒性分析

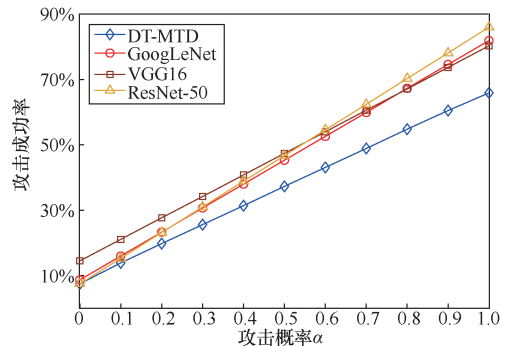
对于 4.1 节中防御者和用户的收益, 可以通过求解式(11)得到防御者切换成员模型的策略和系统分类准确率。为了验证 DT-MTD 的防御对抗样本的有效性, 在攻击者出现的先验概率 α 下, 将对抗样本攻击 DT-MTD 和单个成员模型的攻击成功率进行对比, 实验结果如图 1 所示, 其中(a)、(b)、(c)和(d)分别对应攻击者使用 FGSM、PGD、MI-FGSM 和 M-DI²-FGSM 生成的对抗样本攻击。

以图 1(b)为例, 分以下几种情况进行分析。

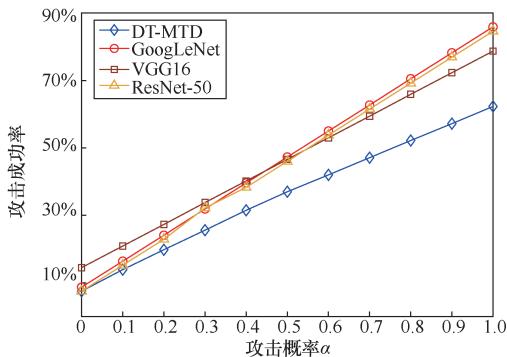
① 假设用户类型只有合法用户, 此时设置 $\alpha = 0$, 图像分类系统会选择准确率最高的成员模型 ResNet-50, 相当于 DT-MTD 使用纯策略 ResNet-50, 不启用模型切换。② 假设用户类型



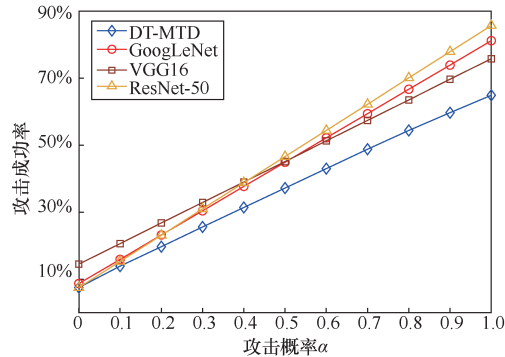
(a) FGSM 的攻击成功率与攻击概率的关系



(b) PGD 的攻击成功率与攻击概率的关系



(c) MI-FGSM 的攻击成功率与攻击概率的关系



(d) M-DI²-FGSM 的攻击成功率与攻击概率的关系

图 1 不同的攻击者在不同概率下, 使用 DT-MTD 的图像分类系统与单个成员模型的攻击成功率对比
Figure 1 Comparison of attack success rate of DT-MTD classification system with different attackers probability

只有攻击者,即所有的请求都是对抗样本,此时设置 $\alpha=1$,DT-MTD求解得到的最优调度策略是 $s=(0.27,0.59,0.14)$,DT-MTD根据概率向量 s 随机选择成员模型。在此策略下,对抗样本对于图像分类系统的攻击成功率为65.9%,但对于大多数单个成员模型,对抗样本的攻击成功率大于80%,特别是对模型ResNet-50,对抗样本的攻击成功率为86%。③实际情况通常是合法用户和攻击者以一定的先验概率分布,假设攻击者出现的概率为 $\alpha(0<\alpha<1)$,那么合法用户的出现概率就为 $1-\alpha$ 。实验中, α 分别取0.1,0.2,...,0.9,模拟各种可能的情况。从图1中可以观察发现,随着 α 的不断增大,即对抗样本在请求中的比例增加,所有模型的攻击成功率都出现上升趋势,这是由于对抗样本的破坏效果在增加。但可以发现,对抗样本对于DT-MTD的攻击成功率仍低于单一成员模型。对于使用FGSM、MI-FGSM和M-DI²-FGSM的对抗攻击,DT-MTD攻击成功率同样低于单一成员模型,由此可见,本文提出的DT-MTD方法可以有效提高图像分类系统的鲁棒性。

4.3 参数 λ 对DT-MTD性能的影响

为了分析差异性训练中的参数 λ 对DT-MTD性能的影响,本文假设最坏的情况下,即用户全部为攻击者($\alpha=1$),对于不同的攻击者,防御者的分类准确率与参数 λ 的关系如图2所示。以攻击者使用PGD生成对抗样本为例,当DT-MTD的成员模型没有进行差异性训练,即 $\lambda=0$ 时,此时DT-MTD相当于Sengupta等^[7]提出的MTD方法,图像分类系统对攻击者使用PGD生成的对抗样本的分类准确率是16.3%。随着 λ 值的不断增大到1时,图像分类系统分类对抗样本的准确率提高到46.7%。对于FGSM、MI-FGSM和M-DI²-FGSM生成的对抗样本也具有相同的趋势,该结果说明了本文提出的差异性训练方法使MTD更加有效。

表6表示不同的差异性训练参数 λ ,对图像分类系统差分免疫 DI 的影响。以使用PGD生成对抗样本的攻击者为例,当差异性训练的参数 $\lambda=0$,差分免疫 DI 值是0.18,随着 λ 的不断增大,差分免疫 DI 值不断增大,在 $\lambda=1$ 时

$DI=0.52$,这与文中3.1节分析的结果一致,差异性训练增大了模型之间的差异性,使图像分类系统具有更大的差分免疫。

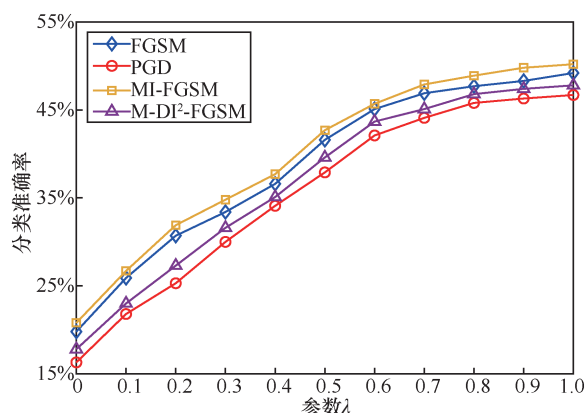


图2 防御者的分类准确率 λ 与参数的关系

Figure 2 Relationship between defender's classification accuracy and parameter λ

表6 差异性训练参数 λ 值(0,0.2,...,1)对图像分类系统差分免疫 DI 的影响

Table 6 Impact of the differential adversarial training parameter λ on differential immune DI

对抗样本生成方法	λ					
	0	0.2	0.4	0.6	0.8	1
FGSM	0.19	0.26	0.34	0.41	0.46	0.48
PGD	0.18	0.28	0.39	0.452	0.5	0.52
MI-FGSM	0.19	0.29	0.42	0.48	0.53	0.55
M-DI ² -FGSM	0.17	0.25	0.37	0.48	0.55	0.59

5 结束语

本文首先介绍了深度学习安全问题领域的相关内容,探讨了图像分类系统基于Bayes-Stackelberg博弈的移动目标防御,通过分析图像分类系统的差分免疫,提出了差异性训练成员模型方法。实验表明,相比于成员模型未进行任何差异性处理,本文提出的差异性训练可以显著提高DNN的移动目标防御性能。未来的工作中,计划将移动目标防御与已有的防御技术相结合,进一步探索深度学习模型的安全性研究。

参考文献:

- [1] HOLI G, JAIN D K. Convolutional neural network approach for extraction and recognition of digits from bank cheque images[M]//Emerging Research in Electronics, Computer Science and

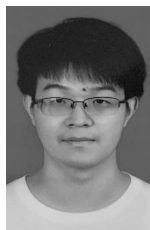
Technology. 2019: 331-341.

- [2] FOMIN I S, BAKHSHEV A V. Research on convolutional neural network for object classification in outdoor video surveillance system[C]//International Conference on Neuroinformatics. 2019: 221-229.
- [3] BUYVAL A, GABDULLIN A, LYUBIMOV M. Road sign detection and localization based on camera and lidar data[C]//Eleventh International Conference on Machine Vision (ICMV 2018). 2019: 1104125.
- [4] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[J]. arXiv preprint arXiv:1312.6199, 2013.
- [5] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[J]. arXiv preprint arXiv:1412.6572, 2014.
- [6] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks[J]. arXiv preprint arXiv:1706.06083, 2017.
- [7] SENGUPTA S, CHAKRABORTI T, KAMBHAMPATI S. MTDeep: moving target defense to boost the security of deep neural nets against adversarial attacks[M]. Proc GameSec, 2019.
- [8] ZHUANG R, DELOACH S A, OU X. Towards a theory of moving target defense[C]//Proceedings of the First ACM Workshop on Moving Target Defense. 2014: 31-40.
- [9] RUSSAKOVSKY O, DENG J, SU H, et al. ImageNet large scale visual recognition challenge[J]. International Journal of Computer Vision, 2015, 115(3): 211-252.
- [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] 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.
- [12] PARUCHURI P, PEARCE J P, TAMBE M, et al. An efficient heuristic approach for security against multiple adversaries[C]//In AAMAS. 2007.
- [13] PARUCHURI P, PEARCE J P, MARECKI J, et al. Playing games for security: an efficient exact algorithm for solving Bayesian-Stackelberg games[C]//Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems- Volume 2. 2008: 895-902.
- [14] DENG J, DONG W, SOCHER R, et al. ImageNet: a large-scale hierarchical image database[C]//IEEE Conference on Computer Vision & Pattern Recognition. IEEE, 2009.
- [15] 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.
- [16] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.
- [17] 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.

[作者简介]



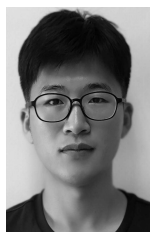
王滨（1978—），男，山东泗水人，海康威视数字技术有限公司研究员，主要研究方向为人工智能安全、物联网安全、密码学等。



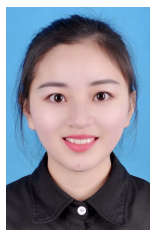
陈靓（1995—），男，江苏无锡人，浙江科技学院硕士生，主要研究方向为对抗深度学习、神经网络压缩。



钱亚冠（1976—），男，浙江嵊州人，博士，浙江科技学院副教授，主要研究方向为人工智能安全、机器学习与大数据处理、对抗性机器学习。



郭艳凯（1994—），男，河南驻马店人，浙江科技学院硕士生，主要研究方向为深度学习图像处理、对抗深度学习。



邵琦琦（1997—），女，浙江永嘉人，浙江科技学院硕士生，主要研究方向为深度学习安全。



王佳敏（1993—），女，浙江新沂人，浙江科技学院硕士生，主要研究方向为深度学习安全。