



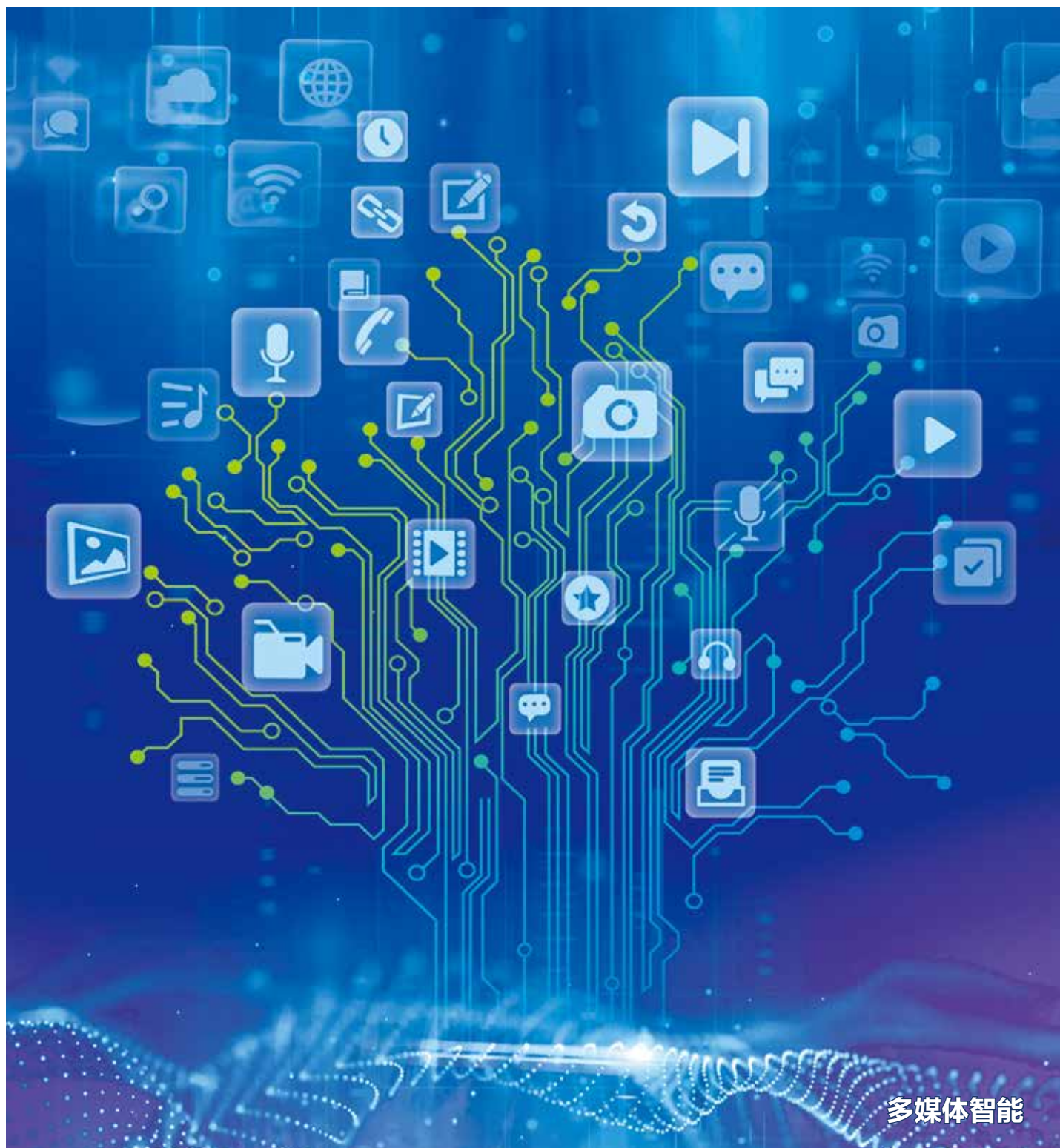
JOURNAL OF IMAGE AND GRAPHICS

主办: 中国科学院空天信息创新研究院
中国图象图形学学会
北京应用物理与计算数学研究所

中国图象图形学报

2022
09
VOL.27

ISSN1006-8961
CN11-3758/TB



多媒体智能



第27卷第9期（总第317期）
2022年9月16日

中国精品科技期刊
中国国际影响力优秀学术期刊
中国科技核心期刊
中文核心期刊

版权声明

凡向《中国图象图形学报》投稿，均视为同意在本刊网站及CNKI等全文数据库出版，所刊载论文已获得著作权人的授权。本刊所有图片均为非商业目的使用，所有内容，未经许可，不得转载或以其他方式使用。

Copyright

All rights reserved by Journal of Image and Graphics, Institute of Remote Sensing and Digital Earth, CAS. The content (including but not limited text, photo, etc) published in this journal is for non-commercial use.

主管单位 中国科学院
主办单位 中国科学院空天信息创新研究院
中国图象图形学学会
北京应用物理与计算数学研究所

主 编 吴一戎
编辑出版 《中国图象图形学报》编辑出版委员会
通信地址 北京市海淀区北四环西路19号
邮 编 100190
电子信箱 jig@aircas.ac.cn
电 话 010-58887035
网 址 www.cjig.cn

广告发布登记号 京朝工商广登字20170218号
总 发 行 北京报刊发行局
订 购 全国各地邮局
海外发行 中国国际图书贸易集团有限公司
(邮政信箱: 北京399信箱 邮编: 100048)
印刷装订 北京科信印刷有限公司

Journal of Image and Graphics

Title inscription: Song Jian | Monthly, Started in 1996

Superintended by Chinese Academy of Sciences
Sponsored by Aerospace Information Research Institute, CAS
China Society of Image and Graphics
Institute of Applied Physics and Computational Mathematics

Editor-in-Chief Wu Yirong
Editor, Publisher Editorial and Publishing Board of Journal of Image and Graphics
Address No. 19, North 4th Ring Road West, Haidian District, Beijing, P. R. China
Zip code 100190
E-mail jig@aircas.ac.cn
Telephone 010-58887035
Website www.cjig.cn

Distributed by Beijing Bureau for Distribution of Newspapers and Journals
Domestic All Local Post Offices in China
Overseas China International Book Trading Corporation
(P.O.Box 399, Beijing 100048, P.R.China)
Printed by Beijing Kexin Printing Co., Ltd.

CN 11-3758/TB
ISSN 1006-8961
CODEN ZTTXFZ

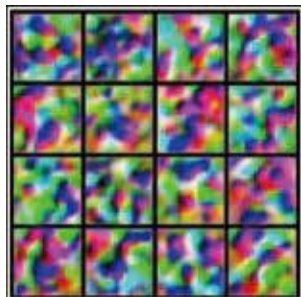
国外发行代号 M1406
国内邮发代号 82-831
国内定价 60.00元



多媒体智能：当多媒体遇到人工智能(第2551页)



面向海洋的多模态智能计算：挑战、进展和展望(第2589页)



基于真实数据感知的模型功能窃取攻击(第2721页)

《中国图象图形学报》多媒体智能专刊简介

朱文武, 黄庆明, 黄华, 蒋树强, 彭宇新, 刘青山, 王井东, 纪荣嵘, 邓伟洪, 方玉明, 刘家瑛, 韩向娣 2549

学者观点

多媒体智能：当多媒体遇到人工智能

朱文武, 王鑫, 田永鸿, 高文 2551

视觉知识：跨媒体智能进化的新支点

杨易, 庄越挺, 潘云鹤 2574

面向海洋的多模态智能计算：挑战、进展和展望

聂婕, 左子杰, 黄磊, 王志刚, 孙正雅, 仲国强, 王鑫, 王玉成, 刘安安, 张弘, 董军宇, 魏志强 2589

综述

基于深度学习的人—物交互关系检测综述

廖越, 李智敏, 刘侃 2611

人类面部重演方法综述

刘锦, 陈鹏, 王茜, 付晓蒙, 戴娇, 韩冀中 2629

视觉语言多模态预训练综述

张浩宇, 王天保, 李孟择, 赵洲, 浦世亮, 吴飞 2652

Bayer阵列图像去马赛克算法综述

魏凌云, 孙帮勇 2683

多媒体智能安全

多特征决策融合的音频copy-move篡改检测与定位

张国富, 肖锐, 苏兆品, 廉晨思, 岳峰 2697

多级特征全局一致性的伪造人脸检测

杨少聪, 王健, 孙运莲, 唐金辉 2708

基于真实数据感知的模型功能窃取攻击

李延铭, 李长升, 余佳奇, 袁野, 王国仁 2721

目标智能检测

利用时空特征编码的单目标跟踪网络

王蒙蒙, 杨小倩, 刘勇 2733

结合时空一致性的FairMOT跟踪算法优化

彭嘉淇, 王涛, 陈柯安, 林巍峭 2749

多媒体分析与理解

融合知识表征的多模态Transformer场景文本视觉问答方法

余宙, 俞俊, 朱俊杰, 匡振中 2761

结合多层级解码器和动态融合机制的图像描述

姜文晖, 占锟, 程一波, 夏雪, 方玉明 2775

面向非受控场景的人脸图像正面化重建

辛经纬, 魏子凯, 王楠楠, 李洁, 高新波 2788

CONTENTS

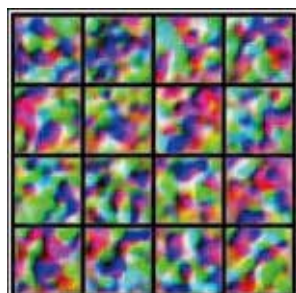
JOURNAL OF IMAGE AND GRAPHICS



Multimedia intelligence: the convergence of multimedia and artificial intelligence(P2551)



Marine oriented multimodal intelligent computing: challenges, progress and prospects(P2589)



Model functionality stealing attacks based on real data awareness(P2721)

Scholar View

Multimedia intelligence: the convergence of multimedia and artificial intelligence

Zhu Wenwu, Wang Xin, Tian Yonghong, Gao Wen 2551

The review of visual knowledge: a new pivot for cross-media intelligence evolution

Yang Yi, Zhuang Yueting, Pan Yunhe 2574

Marine oriented multimodal intelligent computing: challenges, progress and prospects

Nie Jie, Zuo Zijie, Huang Lei, Wang Zhigang, Sun Zhengya, Zhong Guoqiang, Wang Xin, Wang Yucheng, Liu An'an, Zhang Hong, Dong Junyu, Wei Zhiqiang 2589

Review

A review of deep learning based human-object interaction detection

Liao Yue, Li Zhimin, Liu Si 2611

Critical review of human face reenactment methods

Liu Jin, Chen Peng, Wang Xi, Fu Xiaomeng, Dai Jiao, Han Jizhong 2629

Comprehensive review of visual-language-oriented multimodal pre-training methods

Zhang Haoyu, Wang Tianbao, Li Mengze, Zhao Zhou, Pu Shiliang, Wu Fei 2652

The review of demosaicing methods for Bayer color filter array image

Wei Lingyun, Sun Bangyong 2683

Multimedia Intelligent Security

Multi-feature decision fused detection and localization method for copy-move forgery of digital audio clips

Zhang Guofu, Xiao Rui, Su Zhaopin, Lian Chensi, Yue Feng 2697

Multi-level features global consistency for human facial deepfake detection

Yang Shaocong, Wang Jian, Sun Yunlian, Tang Jinhui 2708

Model functionality stealing attacks based on real data awareness

Li Yanming, Li Changsheng, Yu Jiaqi, Yuan Ye, Wang Guoren 2721

Object Intelligent Detection

A spatio-temporal encoded network for single object tracking

Wang Mengmeng, Yang Xiaoqian, Liu Yong 2733

Spatio-temporal consistency based FairMOT tracking algorithm optimization

Peng Jiaqi, Wang Tao, Chen Kean, Lin Weiyao 2749

Multimedia Analysis and Understanding

Knowledge-representation-enhanced multimodal Transformer for scene text visual question answering

Yu Zhou, Yu Jun, Zhu Junjie, Kuang Zhenzhong 2761

The integrated mechanism of hierarchical decoders and dynamic fusion for image captioning

Jiang Wenhui, Zhan Kun, Cheng Yibo, Xia Xue, Fang Yuming 2775

Face frontalization for uncontrolled scenes

Xin Jingwei, Wei Zikai, Wang Nannan, Li Jie, Gao Xinbo 2788

中图法分类号: TP389.1 文献标识码: A 文章编号: 1006-8961(2022)09-2721-12

论文引用格式: Li Y M, Li C S, Yu J Q, Yuan Y and Wang G R. 2022. Model functionality stealing attacks based on real data awareness. Journal of Image and Graphics, 27(09): 2721-2732 [李延铭, 李长升, 余佳奇, 袁野, 王国仁. 2022. 基于真实数据感知的模型功能窃取攻击. 中国图象图形学报, 27(09): 2721-2732] [DOI: 10.11834/jig.211265]

基于真实数据感知的模型功能窃取攻击

李延铭, 李长升*, 余佳奇, 袁野, 王国仁

北京理工大学计算机学院, 北京 100081

摘要: **目的** 模型功能窃取攻击是人工智能安全领域的核心问题之一, 目的是利用有限的与目标模型有关的信息训练出性能接近的克隆模型, 从而实现模型的功能窃取。针对此类问题, 一类经典的工作是基于生成模型的方法, 这类方法利用生成器生成的图像作为查询数据, 在同一查询数据下对两个模型预测结果的一致性进行约束, 从而进行模型学习。然而此类方法生成器生成的数据常常是人眼不可辨识的图像, 不含有任何语义信息, 导致目标模型的输出缺乏有效指导性。针对上述问题, 提出一种新的模型窃取攻击方法, 实现对图像分类器的有效功能窃取。**方法** 借助真实的图像数据, 利用生成对抗网络 (generative adversarial net, GAN) 使生成器生成的数据接近真实图像, 加强目标模型输出的物理意义。同时, 为了提高克隆模型的性能, 基于对比学习的思想, 提出一种新的损失函数进行网络优化学习。**结果** 在两个公开数据集 CIFAR-10 (Canadian Institute for Advanced Research-10) 和 SVHN (street view house numbers) 的实验结果表明, 本文方法能够取得良好的功能窃取效果。在 CIFAR-10 数据集上, 相比目前较先进的方法, 本文方法的窃取精度提高了 5%。同时, 在相同的查询代价下, 本文方法能够取得更好的窃取效果, 有效降低了查询目标模型的成本。**结论** 本文提出的模型窃取攻击方法, 从数据真实性的角度出发, 有效提高了针对图像分类器的模型功能窃取攻击效果, 在一定程度上降低了查询目标模型代价。

关键词: 模型功能窃取; 生成模型; 对比学习; 对抗攻击; 人工智能安全

Model functionality stealing attacks based on real data awareness

Li Yanming, Li Changsheng*, Yu Jiaqi, Yuan Ye, Wang Guoren

School of Computer Science, Beijing Institute of Technology, Beijing 100081, China

Abstract: **Objective** Current model stealing attack issue is a sub-field in artificial intelligence (AI) security. It tends to steal privacy information of the target model including its structures, parameters and functionality. Our research is focused on the model functionality stealing attacks. We target a deep learning based multi-classifier model and train a clone model to replicate the functionality of the black-box target classifier. Currently, most of stealing-functionality-attacks are oriented on querying data. These methods replicate the black-box target classifier by analyzing the querying data and the response from the target model. The kind of attacks based on generative models is popular and these methods have obtained promising results in functionality stealing. However, there are two main challenges to be faced as mentioned below: first, target image classifiers are trained on real images in common. Since these methods do not use ground truth data to supervise the training phase of generative models, the generated images are distorted to noise images rather than real images. In other words, the image data used by these methods is with few semantic information, leading to that the prediction of target model is with few effective guidance for the training of the clone model. Such images restrict the effect of training the clone model.

收稿日期: 2022-01-11; 修回日期: 2022-04-28; 预印本日期: 2022-05-05

* 通信作者: 李长升 lcs@bit.edu.cn

Second, to train the generative model, it is necessary to initiate multiple queries to the target classifier. A severe burden is bear on query budgets. Since the target model is a black-box model, we need to use its approximated gradient to obtain generator via zero-gradient estimation. Hence, the generator cannot obtain accurate gradient information for updating itself.

Method We try to utilize the generative adversarial nets (GAN) and the contrastive learning to steal target classifier functionality. The key aspect of our research is on the basis of the GAN-based prior information extraction of ground truth images on public datasets, aiming to make the prediction from the target classifier model be with effective guidance for the training of the clone model. To make the generated images more realistic, the public datasets are introduced to supervise the training of the generator. To enhance the effectiveness of generative models, we adopt deep convolutional GAN (DCGAN) as the backbone, where the generator and discriminator are composed of convolutional layers both with non-linear activation functions. To update the generator, we illustrate the target model derived gradient information via zero-order gradient evaluation for the training of clone model. Simultaneously, we leverage the public dataset to guide the training of the GAN, aiming to make the generator obtain the information of ground truth images. In other words, the public dataset plays a role as a regularization term. Its application constrains the solution space for the generator. In this way, the generator can produce approximated ground truth images to make the prediction of the target model produce more physical information for manipulating the clone model training. To reduce the query budgets, we pre-train the GAN on public datasets to make it obtain prior information of real images before training the clone model. Our method can make the generator learn better for the training need of clone model in comparison with previous approaches of the random-initialized generator training. To expand the objective function of training clone model, we introduce contrastive learning to the model stealing attacks area. Traditional model functionality stealing attack methods train the clone model only by maximizing the similarity of predictions from two models to one image. Here, we use the contrastive learning manner to consider the diversity of predictions from two models to different images. The positive pair consists of the predictions from two models to one image and the negative pair is made up with the predictions from two models to two different images. To measure the diversity of two predictions, we attempt to use cosine similarity to represent the similarity of two predictions. Then, we use the InfoNCE loss function to achieve the similarity maximization of positive pairs and diversity maximization of negative pairs at the same time.

Result To demonstrate the performances of our methods, we carry out model functionality stealing attacks on two different black-box target classifiers. The two classifiers of Canadian Institute for Advanced Research-10 (CIFAR-10) and street view house numbers (SVHN) are presented. Each of model structure is based on ResNet-34 and the structures of clone models are based on resnet-18 both. The used public datasets are not be overlapped with the training datasets of target classifiers. We test them on CIFAR-10 and SVHN test datasets following our trained clone models. The accuracy results of these clone models are 92.3% and 91.8% of each. Normalized clone accuracy is achieved $0.97 \times$ and $0.98 \times$ of each. Specially, our result can achieve 5% improvements for the CIFAR-10 target model in terms of normalized clone accuracy over the data-free model extraction (DFME). Our method achieves promising results for reducing querying budgets as well. To make the accuracy of clone model reach 85% on the CIFAR-10 test datasets, DFME is required to spend 8.6 M budgets. But, our method spends 5.8 M budgets only, which is 2.8 M smaller than DFME. Our method is required to spend 9.4 M budgets for reaching the 90% accuracy, which is not half enough to the DFME of 20 M budgets. These results demonstrate that our method improve the performances of functionality stealing attack methods based on generative models. It is beneficial for reducing the query budgets as well.

Conclusion We propose a novel model functionality stealing attack method, which trains the clone model guided by prior information of ground truth images and the contrastive learning manner. The experimental results show that our optimized model has its potentials and the querying budgets can be reduced effectively.

Key words: model functionality stealing; generative model; contrastive learning; adversarial attack; artificial intelligence security

0 引言

随着机器学习尤其是深度学习在计算机视觉和

自然语言处理等任务中大放异彩,企业界开始尝试通过机器学习或深度学习为用户提供相关的智能服务。机器学习模型在自动驾驶、语音识别、图像分类和医疗图像分析(黎英和宋佩华,2022)等领域已得

到广泛应用。众所周知,训练深度学习模型需要大量训练数据(Strubell等,2019),而收集这些训练数据需要耗费大量时间成本和经济成本,设计深度学习模型或架构需要额外的专家知识(Brown等,2020)。因此,拥有高性能的深度学习模型通常视为拥有昂贵的知识产权。基于此,商业公司通过向用户开放API(application programming interface)接口来提供服务,即用户向模型提交查询数据,获得该模型的预测结果,而公司通过收取查询费用实现收益。在整个过程中深度学习模型对用户来说是黑盒模型,但是模型的预测结果中携带了模型的部分信息,使得攻击者可以利用这些信息实现对该模型功能的窃取(Tramèr等,2016)。

当前,模型功能窃取攻击的主流框架如图1所示。以图像分类器为例,攻击者首先通过提交查询图像到目标模型(即要攻击的模型)获得预测结果;随后利用查询图像和预测结果训练一个克隆模型。通过此方式使克隆模型获得与目标模型相似的功能。该方法与知识蒸馏技术(Hinton等,2015)十分相似,克隆模型类似于知识蒸馏中的学生模型(Fang等,2020),优化目标是在面对相同数据时使克隆模型的行为能尽量与目标模型(对应知识蒸馏中的教师网络)一致。与知识蒸馏不同的是,由于目标模型对攻击者来说是一个黑盒模型,因此训练数据、内部结构和优化器设计等信息对攻击者来说都是不可知的。对攻击者来说,假设可知的信息仅有其提交的查询数据和该查询数据经过目标模型返回的预测结果。不同查询数据集常常导致克隆模型获得的性能差异较大。目前已经提出了很多模型功能窃取方法。Knockoff(Orekondy等,2019)是模型窃取攻击方法中的一个经典工作,目的是利用一些公开数据集对目标模型发起查询请求进而构造数据对集合 $D = \{(\mathbf{x}_1, \mathbf{y}_1), \dots, (\mathbf{x}_n, \mathbf{y}_n)\}$ 训练克隆模型,其中 $\mathbf{x}_i$ 和 $\mathbf{y}_i$ 分别为图像数据及目标模型对该图像的预测结果。此外,学者相继提出了一些无数据的模型窃取方法。DFME(data-free model extraction)(Truong等,2021)在没有任何真实数据的情况下,利用零阶梯度估计(Chen等,2017)训练生成模型生成数据,且利用数据对集合 D 训练克隆模型。然而由于缺乏真实图像数据的介入,这类方法生成的数据在人眼看来是不可分辨的,不包含有效的语义信息,使得目标模型输出的物理意义不明确,从而造成

对克隆模型的训练缺乏有效的指导。

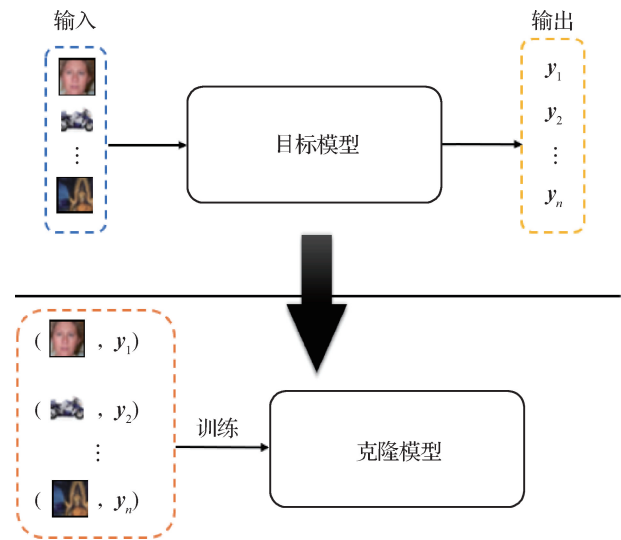


图1 模型功能窃取攻击实例

Fig. 1 An example for the model stealing attack

本文从加强目标模型输出的物理意义出发,通过真实数据感知的方式,提出一种新的模型窃取攻击方法,实现对图像分类器的有效功能窃取。针对上述方法在生成数据时缺乏语义信息问题,引入真实数据,使生成器在生成图像时不仅要考虑满足模型功能窃取任务的需要,而且要考虑生成图像的真实性,进而加强目标模型输出结果对克隆模型训练指导的有效性。具体来讲,引入公开的真实数据对生成对抗网络(generative adversarial network, GAN)(Goodfellow等,2014)的学习进行约束,通过令Discriminator分类器区分数据是来自真实数据集还是生成数据,在如此博弈方式下使生成器生成的数据集与真实数据集分布相似,从而克服数据不可辨和语义信息缺失问题,加强了目标模型输出结果的物理意义。此外,提出一种新的目标函数对克隆模型进行优化学习,该目标函数除了对两个模型预测结果做一致性约束外,引入对比学习(Hadsell等,2006)思想,在考虑两个模型对同一图像数据分类结果一致性的同时,将不同图像预测结果的差异性也考虑进来。在两个公开图像数据集的分类模型上,本文方法在窃取效果和查询代价上都取得了很好效果。本文主要贡献如下:1)首次从生成数据的真实性出发,利用生成对抗网络生成图像数据,使生成图像的分布与真实图像的分布相似,提高模型输出的物理意义,加强目标模型的有效指导;2)提出一种新的目标函数,基于一致性输出和对比学习思

想,对目标模型和克隆模型的输出进行约束,达到训练有效克隆模型的目的;3)实验结果表明,本文方法在对图像分类器模型进行功能窃取时,能够取得良好效果,并且在一定程度上降低了对目标模型的查询代价。

1 相关工作

模型窃取攻击是人工智能安全中的核心问题之一,目的是推断目标黑盒模型的隐私信息,包括内部结构(Oh等,2019)、参数(Wang和Gong,2018)和预测功能(Correia-Silva等,2018)等。本文主要研究模型的功能窃取。由于此场景下目标模型为黑盒模型,其内部结构信息、模型参数和训练数据等对攻击者来说均是不可知的,因此攻击方法常常通过分析目标模型的输入与输出的映射关系,训练一个与目标模型功能相似的克隆模型。经典方法包括基于查询数据设计的方法(Gong等,2021)和基于损失函数设计的方法(Truong等,2021)。

1.1 模型窃取攻击

根据攻击者使用的训练数据不同,目前的模型功能窃取攻击大致分为3类(Kariyappa等,2021),分别是部分训练数据已知、有可用的替代数据集以及利用生成模型生成数据的方法。

在部分训练数据已知情况下,已知数据的数据规模较小,仅使用这些数据无法训练出令人满意的克隆模型,攻击者会利用这些已知的训练数据生成训练样本来增大数据量。JBDA(Jacobian based dataset augmentation)(Papernot等,2017)是此类工作中的经典方法,其利用雅可比矩阵来表示模型对输入数据的敏感程度,并且在雅可比矩阵的基础上指导对抗样本的生成,最后利用这些对抗样本和已知的训练数据训练克隆模型。

在攻击者有可用的替代数据集情况下,攻击者利用该数据集访问目标模型,并将目标模型对该数据集的预测结果当作对应数据的标签(或软标签),然后训练克隆模型。Knockoff(Orekondy等,2019)就是在这一情景下提出的。当有合适的替代数据集时,该窃取方法会取得较好的效果。然而对于攻击者来说,目标模型所对应训练数据的信息是未知的,因此在实际场景中攻击者很难选择一个合适的替代数据集。

在利用生成模型生成训练数据的方法中,攻击者没有任何可用的真实数据信息,其用于对目标模型发起查询的数据均是通过生成模型生成的。由于没有真实数据,相较于前两种情况,这种情况下模型功能窃取的任务更加困难。MAZE(data-free model stealing attack using zeroth-order gradient estimation)(Kariyappa等,2021)和DFME(data-free model extraction)(Truong等,2021)是两个经典的生成模型的方法,其思想是利用零阶梯度估计的方法(Duchi等,2012)对黑盒模型的梯度进行近似,以对抗博弈的思想更新生成器的参数和克隆模型的学习。尽管这类方法取得了良好效果,但是由于生成器的训练需要与目标模型进行额外的交互,因此具有较高的查询复杂度。

1.2 生成对抗网络

生成对抗网络(GAN)(Goodfellow等,2014)是目前最受欢迎的生成模型之一。与传统生成模型不同,该方法通过一种对抗训练的策略来规避对数据概率分布的直接求解。GAN网络由生成器(generator)和判别器(discriminator)两部分组成。生成器的训练目标是生成尽量真实的数据来迷惑判别器;而判别器的目标是能够区分哪些样本是真实的,哪些样本是生成器生成的,通过如此方式使生成器生成的数据与真实数据具有相似的分布。随着GAN网络的广泛流行,提出了越来越多的扩展性工作。例如,针对复杂图像数据生成的DCGAN(deep convolutional generative adversarial network)(Radford等,2016)和用来做风格迁移任务的StyleGAN(Karras等,2019)。也有人开始尝试利用GAN提升取证技术的效果(何沛松等,2022)。在模型功能窃取方法中,一些工作开始尝试借鉴GAN的思想在无数据情况下通过生成模型生成数据来对目标模型进行查询。例如DFME和MAZE虽然没有显式引入类似于GAN中的判别器,但目标模型和克隆模型本身起到了判别器的作用。

1.3 对比学习

对比学习(Hadsell等,2006)是比较流行的一种自监督学习方法,核心思想是在缩小正样本对之间距离的同时拉大负样本对之间的距离,已广泛应用于各领域,例如表示学习。目前,对比学习的方法有许多。MoCo(momentum contrast for unsupervised visual representation learning)(He等,2020)通过设计

动态队列和非对称更新的编码器网络在对比学习的指导下学习数据的表征,使得在 ImageNet (Deng 等, 2009) 上经过无监督预训练的模型可以很好地泛化到一些下游任务。SimCLR (simple framework for contrastive learning of visual representations) (Chen 等, 2020) 使用较大的 batch 增大对比学习中负样本对的数量,同样达到了提升表示学习效果的目的。本文将对比学习的思想引入模型功能窃取攻击。

2 真实数据感知的模型功能窃取

模型功能窃取的目标是训练一个与目标模型在功能上相近的克隆模型。本文研究要窃取的目标模型为图像分类器,因此克隆模型的训练目标是近似目标模型的分类效果。

2.1 整体方法

先前基于生成模型的方法在生成数据过程中缺少真实数据介入,使生成器生成的数据在人眼看来是不可分辨的,不包含有效的语义信息,导致目标模型输出的物理意义不明确,降低了目标模型指导克隆模型的有效性。

本文从加强目标模型输出具有物理意义的角度出发,提出一种新的模型窃取攻击方法,实现对图像分类器的有效功能窃取。本文方法的核心思想是利

用生成对抗网络 (GAN) 生成图像数据,使生成数据与真实数据集的分布相似,从而加强数据的真实性。本文方法的架构如图 2 所示,首先利用公开数据集对 GAN 网络进行预训练,使其提取真实数据的先验知识,然后通过最小化目标模型和克隆模型输出结果的差异性和 GAN 网络本身的损失函数微调 GAN 网络的参数。在 GAN 网络训练好之后,使用训练好的生成器生成样本数据,用该生成数据作为查询数据分别送入目标模型和克隆模型,通过对两个模型输出结果的一致性进行约束来训练克隆模型。在优化克隆模型时,与之前方法不同的是本文提出了一种新的损失函数,不仅考虑两个模型对同一幅图像分类结果的一致性,同时通过引入对比学习的思想将不同图像预测结果的差异性也考虑进来。

从数据视角出发,本文方法结合了基于替代数据集和无数据两种模型窃取方法的优点,利用 GAN 模型与目标模型的交互,使生成图像在接近公开数据集真实图像分布的同时,又受到目标模型反馈信息的影响,从而在一定程度上能够缓解对公开数据集的过分依赖;同时,由于公开数据集的引入,生成模型可以事先提取到真实数据的先验知识,并且在与目标模型进行交互时,公开数据集在一定程度上约束了生成模型的梯度更新方向,使生成图像更加真实,进而加强了目标模型预测结果的物理意义。

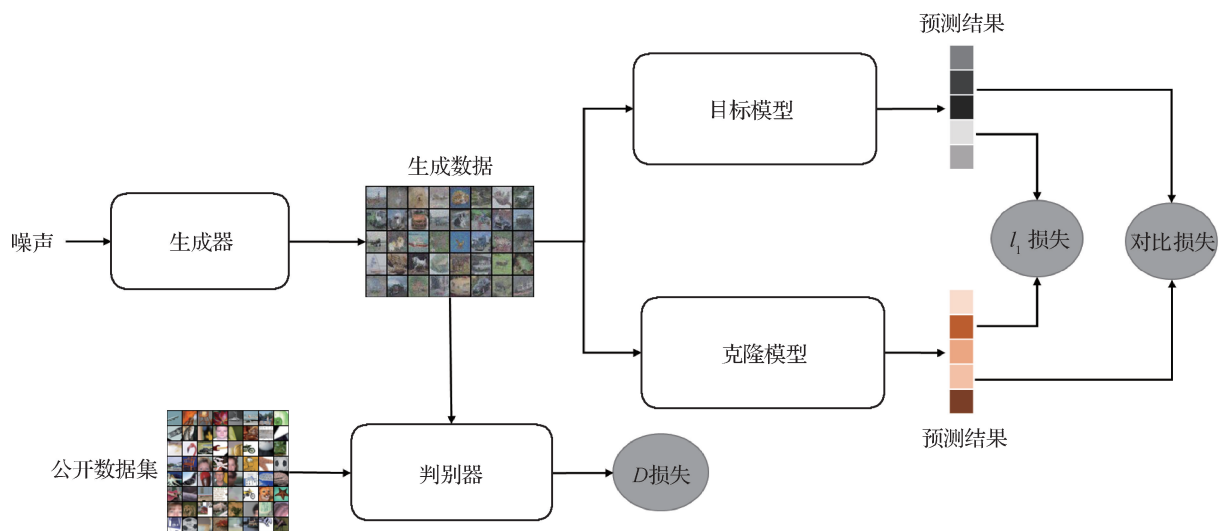


图2 基于真实数据感知的模型窃取攻击架构

Fig. 2 The framework of model stealing attacks based on real data awareness

2.2 生成对抗网络的训练

GAN 网络在整个方法框架中的任务是产生模

型窃取攻击任务依托的数据。考虑到之前方法对生成器的更新仅依靠目标模型和克隆模型反馈的信

息,并且生成器的参数均是随机初始化的,不携带任何先验信息,因此对生成器训练需要目标模型进行大量的查询交互,这导致高额的查询代价。同时由于缺少真实图像数据的监督,生成器生成的图像不携带任何语义信息,导致目标模型的预测结果对克隆模型的指导缺少有效性。

为了使生成器能够快速拟合真实数据的分布,降低其与目标模型的交互查询次数,同时为了使克隆模型能够快速学到目标模型的功能,本文利用公开数据集对 GAN 网络进行预训练,使生成器能够学习真实数据的先验信息,其损失函数定义为

$$\min_G \max_D L_1 = E_{x \sim p_{\text{data}}(x)} [\log D(x)] + E_{z \sim p_z(z)} [\log(1 - D(G(z)))] \quad (1)$$

式中, G, D, z, x 分别是生成器、判别器、随机噪声和图像数据, E 表示数学期望。在预训练结束后,用生成器生成模型功能窃取需要的数据,并通过零阶优化的方法,将近似的梯度信息传递给生成器,使生成器能够生成对克隆模型来说预测困难的图像,这部分的损失函数利用 l_1 范数对两个模型预测结果的 logits 做约束,具体为

$$\max_C L_2 = \sum_{i=1}^N \sum_{j=1}^P |\log(C(G(z_i)))_j - \log(T(G(z_i)))_j| \quad (2)$$

式中, N 和 P 分别表示数据集的大小和预测输出的维度(即类别数), C 和 T 分别表示克隆模型和目标模型。由于目标模型是黑盒模型,本文假设得到的预测结果为分类概率,因此这里用 $\log$ 函数对其目标模型的预测结果进行操作,以近似其 logits。

在使生成器生成对克隆模型来说预测困难的图像的同时,本文同时使生成的图像尽量保持真实性,因此定义 GAN 网络在模型窃取时的损失函数为

$$L_G = L_1 + \lambda_1 L_2 \quad (3)$$

式中, λ_1 为平衡两个损失函数的超参数。

2.3 克隆模型的训练

克隆模型的学习目标是复制目标模型的功能,因此在设计损失函数时,一个很直观的想法是最小化两个模型输出的差异。即

$$\min_C L_3 = \sum_{i=1}^N \sum_{j=1}^P |C(\hat{x}_i)_j - T(\hat{x}_i)_j| \quad (4)$$

式中, $\hat{x}_i$ 指生成器通过随机噪声生成的数据图像,即 $\hat{x}_i = G(z_i)$ 。可以看出,该损失函数仅考虑了单

个样本在目标模型与克隆模型之间的一致性。

为了提升模型窃取的效果,本文还采用了对比学习的思想,将不同样本之间的差异性刻画出来,如图 3 所示。

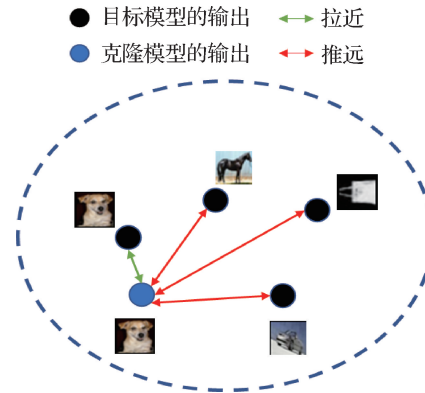


图 3 对比学习的目标

Fig. 3 The goal of the contrastive learning in our method

本文使用 InfoNCE(van den Oord 等,2019)作为对比学习的损失函数。即

$$\min_C L_4 = - \sum_{i=1}^N \log \frac{\exp(\tilde{T}(\hat{x}_i) \cdot \tilde{C}(\hat{x}_i))}{\sum_{j=0}^K \exp(\tilde{T}(\hat{x}_j) \cdot \tilde{C}(\hat{x}_j))} \quad (5)$$

式中, $\hat{x}_j$ 为某个负样本。本文使用一个字典存储目标模型对负样本预测结果,字典大小为 K 。本文用点积操作衡量两个预测结果之间的相似程度,分母上求和的对象是该查询样本在克隆模型上的预测结果与其他样本在目标模型上的预测结果的相似程度。因此通过最小化上面的损失函数,可以拉大不同数据预测结果之间的距离。为了使点积操作能够更好地刻画预测结果之间的相似程度,本文使用归一化的 logits 而非 softmax 层后的分类概率进行点积计算,进而增大点积操作结果的取值范围。与计算 L_2 相同,本文对目标模型的概率输出结果取 $\log$ 操作近似其 logits。即

$$\tilde{T}(\hat{x}_i) = \frac{\log(T(\hat{x}_i))}{\|\log(T(\hat{x}_i))\|_2}, \quad \tilde{C}(\hat{x}_j) = \frac{\log(C(\hat{x}_j))}{\|\log(C(\hat{x}_j))\|_2} \quad (6)$$

因此,在训练克隆模型时,定义总的损失函数为

$$L_C = L_3 + \lambda_2 L_4 \quad (7)$$

式中, λ_2 是为了平衡两个损失函数的超参数。

基于真实数据感知的模型功能窃取的整体算法流程为:

输入:查询代价 Q , GAN 迭代次数 n_{gan} , 克隆模型迭代次数 n_c , 学习率 η , 公开数据集 X , 目标模型 T , 字

典 $dict$ 。

输出: 克隆模型 C 。

while $Q > 0$ do;

 for $i = 1, \dots, n_{\text{gan}}$ do;

$z \sim N(0, I)$;

$x \sim X$ (随机采样);

$\hat{x} = G(z)$;

$d_{\text{real}} = D(x)$;

$d_{\text{fake}} = D(\hat{x})$;

$\text{predict}_t = T(\hat{x})$;

$\text{predict}_c = C(\hat{x})$;

 计算 L_{gan} 并更新 GAN;

 for $i = 1, \dots, n_c$ do;

$z \sim N(0, I)$;

$\hat{x} = G(z)$;

$\text{predict}_t = T(\hat{x})$;

$\text{predict}_c = C(\hat{x})$;

 计算 L_c 并更新克隆模型 C ;

 将 predict_t 存入字典 $dict$ (字典用于计算对比学习损失函数);

 更新查询代价 Q ;

end。

3 实验分析

3.1 数据集和目标模型

为了验证本文方法的有效性,在 SVHN (street view house numbers) (Netzer 等, 2011) 和 CIFAR-10 (Canadian Institute for Advanced Research-10) (Krizhevsky, 2009) 数据集上进行验证。目标模型和克隆模型的结构设定参考 DFME (Truong 等, 2021) 的设置, 分别采用 ResNet-34-8x 和 ResNet-18-8x 作为目标模型和克隆模型的网络结构 (He 等, 2016)。对于 CIFAR-10, 使用 DFME 训练好的目标模型; 对于 SVHN, 使用 Adam 优化器对目标模型进行优化, 训练次数为 150 个 epoch, 学习率为 0.001。当训练完目标模型后将其视为黑盒模型, 即假设模型的内部信息未知; 对于克隆模型, 使用 Adam 优化器对其进行优化, 学习率设置为 0.1。

针对 GAN 网络, 本文采用 DCGAN (deep convolutional generative adversarial network) 的构建方式。生成器由 5 个反卷积层构成, 中间层的激活函数采

用 ReLU 函数, 最后一层采用 Tanh 函数。判别器的网络结构与生成器近似对称, 由 5 个卷积层组成, 中间层的激活函数为 LeakyReLU, 最后一层的激活函数为 sigmoid。这一部分同样用 Adam 优化器进行优化, 在预训练时, GAN 网络生成器和判别器的学习率为 0.000 1。在进行模型窃取时, 对 CIFAR-10 分类模型, 生成器和判别器的学习率均为 5E-5; 对 SVHN 分类模型, 生成器和判别器的学习率均为 0.000 01。整个实验中, batch size 大小保持在 256 不变。在零阶梯度估计中, 仿照 DFME 的设置, 每次随机选取一个方向近似计算梯度。在训练克隆模型时, GAN 网络与克隆模型的训练次数比重为 1:5。本文将算法中的 n_{gan} 设置为 1, n_c 设置为 5, 查询的代价 Q 设置为 20 M。

3.2 总体实验结果

为了充分说明本文方法的效果, 分别从无数据的模型窃取方法和基于替代数据的模型窃取方法中选取基准方法与本文方法进行对比。无数据模型窃取的基准方法包括 DFME (Truong 等, 2021) 和 MAZE (Kariyappa 等, 2021); 基于替代数据集的模型窃取方法包括 Knockoff (Orekondy 等, 2019)。

在测试数据集上对本文方法和基准方法进行测试, 实验结果如表 1 所示。针对 CIFAR-10 和 SVHN 分类模型, 本文使用 Caltech-101 (Li 等, 2004) 作为公开数据集。为了方便对比, 本文参考 Truong 等人 (2021) 的设定, 将查询的代价设置为 20 MB。括号中的数字表示使用目标模型准确率对克隆模型准确率进行归一化的结果 (用 $\times$ 表示), 本文用这一指标量化模型功能窃取的效果。可以看到, 在两个不同的分类器模型上, 本文方法相比于之前较先进的方法在窃取效果上都取得了较大提高。并且本文方法在 CIFAR-10 上的窃取效果达到 0.97 $\times$, 相较于 knockoff 提升了 3%。

表 1 不同模型窃取算法的实验结果

Table 1 The results of different model stealing algorithms

方法	目标模型准确率		克隆模型准确率	
	CIFAR-10	SVHN	CIFAR-10	SVHN
Knockoff	0.955	0.932	0.898 (0.94 $\times$)	0.905 (0.97 $\times$)
MAZE	0.955	0.932	0.456 (0.48 $\times$)	0.859 (0.92 $\times$)
DFME	0.955	0.932	0.881 (0.92 $\times$)	0.917 (0.98 $\times$)
本文	0.955	0.932	0.923 (0.97 $\times$)	0.923 (0.99 $\times$)

注: 加粗字体表示各列最优结果。

3.3 消融实验

3.3.1 不同公开数据集对窃取效果的影响

为验证不同公开数据集对窃取效果的影响,针对 SVHN 目标模型,分析引入不同的公开数据集对窃取效果的影响,包括 Fashion-MNIST (Xiao 等, 2017)、CIFAR-10、MIT Indoor 67 (Quattoni 和 Torralba, 2009) 和 Caltech-101。实验结果如图 4 所示。可以发现,在引入不同的公开数据集时,本文方法针对 SVHN 分类模型功能窃取的效果均超过 DFME,并且在引入的公开数据集(如 Fashion-MNIST)与 SVHN 数据集之间相似性较弱的情况下,本文方法的效果依旧良好,在一定程度上说明了本文方法的有效性。

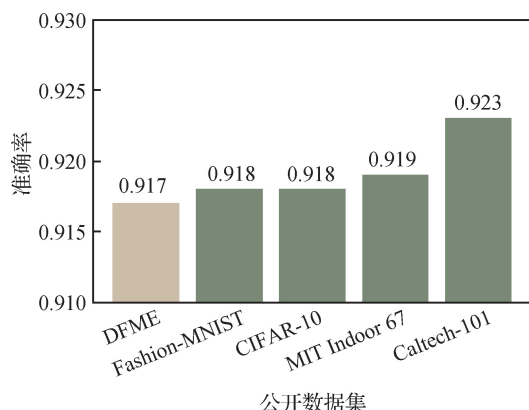


图 4 基于不同公开数据集的功能窃取效果

Fig. 4 The effectiveness of model functionality stealing based on different public datasets

3.3.2 GAN 网络的有效性验证

在模型功能窃取领域,基于替代数据集进行模型窃取是一种经典方法,在针对不同图像分类器时表现较好,但对用作窃取依托的替代数据集有一定限制,要求用来查询的图像和目标模型的训练数据域存在一定相关性,因此这种方法的攻击效果与查询图像和目标模型训练数据间的相似程度成正比。

本文利用 GAN 网络一定程度上缓解了上述问题,以 CIFAR-10 分类器模型为例,采用 3 个公开数据集进行测试,在相同的设定下不使用 GAN 网络进行模型窃取,为了尽可能排除其他因素的干扰,实验均不采用对比学习的损失函数,实验结果如图 5 所示。

结果显示,增加了 GAN 网络的模型实验结果在不同数据集上均明显优于原模型。由于生成器的引入以及生成器与目标模型的交互,该生成器生成的

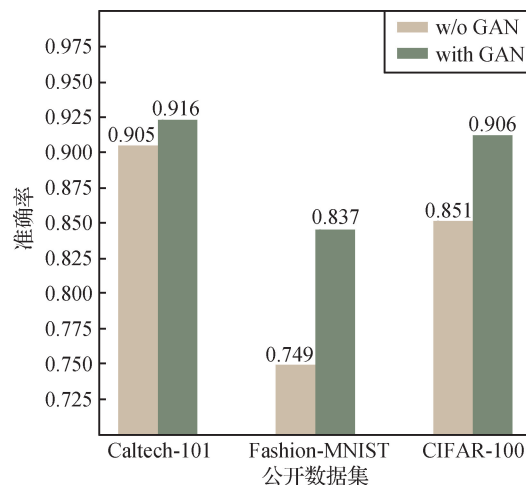


图 5 CIFAR-10 模型在不同公开数据集上有无 GAN 网络的实验效果

Fig. 5 The experimental results of our methods under using or not using GAN for CIFAR-10 target model

样本会携带一些目标模型的先验知识,因此可以在一定程度上缓解对公开数据域与目标模型训练数据域相似程度的依赖。当使用 Fashion-MNIST 数据集对 CIFAR-10 分类器模型进行功能窃取时,克隆模型在 CIFAR-10 测试数据的表现 0.749。引入生成对抗网络之后,克隆模型在 CIFAR-10 测试集上的表现上升到 0.837。可以表明,本文提出的引入 GAN 网络的方法,一定程度上提升了基于替代数据集模型窃取方法的效果。

3.3.3 对比学习损失函数的有效性验证

为验证基于对比学习损失函数的效果进行消融实验,结果如表 2 所示。可以看出,基于对比学习的损失函数不仅对本文方法有效,在 DFME 和 Knock-off 方法上同样会为克隆模型的学习带来收益。同时,在 CIFAR-10 的目标模型上,本文对不同负样本个数进行实验分析,结果如图 6 所示。可以看出,在不同量级的负样本数量下,本文方法的效果波动范围不大,验证了本文方法对负样本个数不敏感。因

表 2 对比学习损失函数测试结果

Table 2 The test results of contrastive loss

方法	不使用对比学习	使用对比学习
Knockoff	0.898 (0.94 ×)	0.918 (0.96 ×)
DFME	0.881 (0.92 ×)	0.892 (0.93 ×)
本文	0.916 (0.96 ×)	0.923 (0.97 ×)

注:括号中结果表示使用目标模型准确率对克隆模型准确率进行归一化的结果。

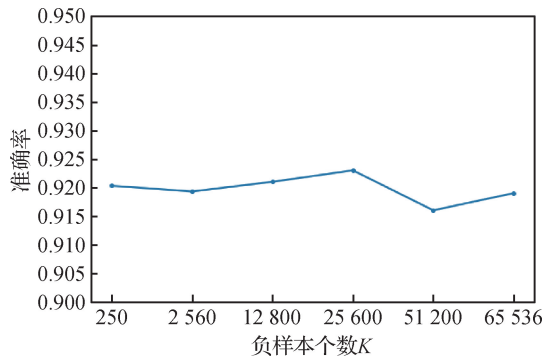


图6 不同负样本数下的实验结果

Fig. 6 The experimental results on different numbers of negative samples

此在实际场景中,负样本的个数较为容易设置。

3.4 查询代价比较

无数据的模型窃取方法虽然可以规避替代数据域与目标模型训练数据域相似程度的限制,但是由于其生成器的训练需要与目标模型进行大量的交互,将会导致查询代价大幅提高。

本文引入公开数据集,可以使生成器提取到真实数据的先验知识,有真实数据监督,生成器可以快速产生有利于模型功能窃取需要的数据,图7和表3展示了本文方法和DFME在CIFAR-10的目标模型上的实验结果。本文方法的测试曲线在同一迭代次数时均高于DFME。以测试结果为85%为例,本文方法需要65个epoch达到此效果,DFME下的克隆模型则需要96个epoch。按照本文的实验设定计算,DFME算法在测试集上达到85%的准确率,相比于本文方法需要额外提交大约2.6 M次请求。以测试结果为90%为例,本文方法下的克隆模型在测试集上达到此效果需要110个epoch,所需查询次数约为9.4 M次,与DFME算法中20 M的查询代价相比,本文方法能够在很大程度上节省查询的代价。

从正则项角度出发,公开数据集在本文方法中的作用可以看成是对生成器的正则项约束。如图8和图9所示,与其他无真实数据监督的模型功能窃取方法对比,DFME由于缺少真实数据监督,生成的数据在人眼看起来没有语义信息,同时仅靠零阶梯度估计方法计算得到的梯度信息更新生成器,使每次生成器的更新带有很大随机性。本文方法在公开数据集的帮助下,生成器的更新伴随着真实数据的指导,利用真实数据约束生成器的更新,对生成器的解空间做了限制,因此本文方法能够更快速地取得良好的功能窃取效果。

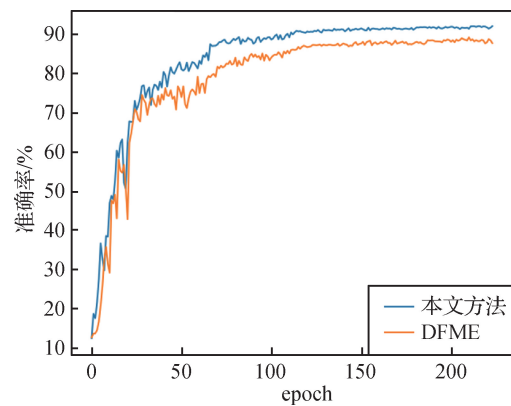


图7 本文方法和DFME在CIFAR-10模型上的测试结果

Fig. 7 The test results of different methods on CIFAR-10 model

表3 在CIFAR-10模型上不同查询代价下的测试准确率

Table 3 The results of different budgets on CIFAR-10 model

查询代价/M	DFME	本文方法
1	0.126	0.124
5	0.767	0.82
8	0.822	0.887
10	0.877	0.906

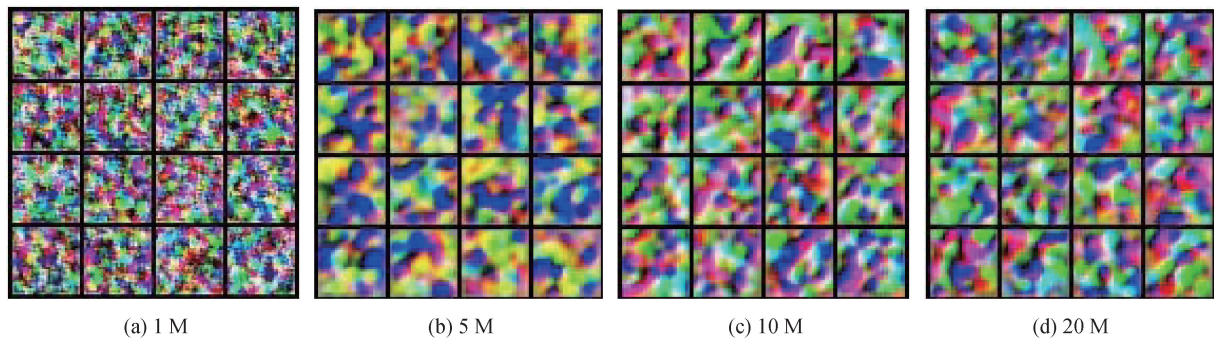


图8 不同代价下DFME方法生成的图像

Fig. 8 The generated images of DFME on different budgets ((a) 1 M; (b) 5 M; (c) 10 M; (d) 20 M)

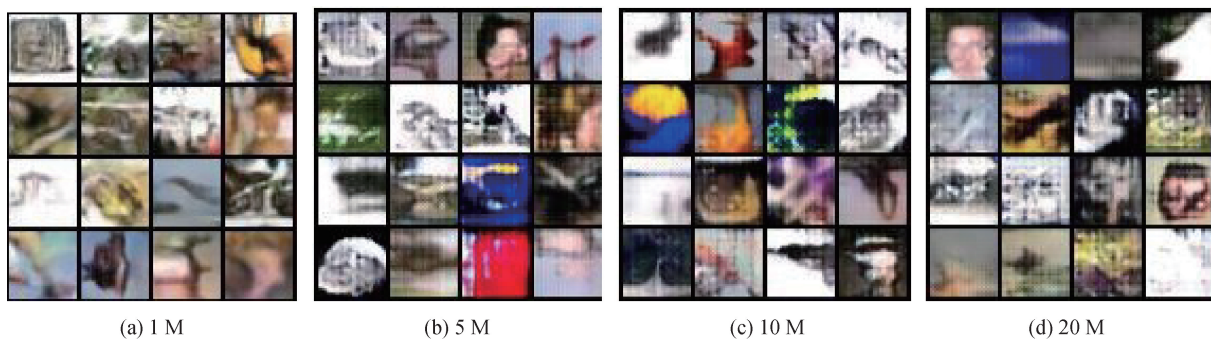


图9 不同代价下本文方法生成的图像

Fig. 9 The generated images of our method on different budgets ((a) 1 M; (b) 5 M; (c) 10 M; (d) 20 M)

4 结 论

本文针对模型功能窃取问题展开研究,从数据真实性的角度出发,引入公开数据集,利用 GAN 网络改进现有基于生成模型进行模型功能窃取方法的缺陷,目的是使生成器生成的数据与真实数据的分布一致。同时,为了提升克隆模型的学习效果,将对对比学习的思想引入到模型功能窃取中。本文方法相比之前最先进的方法在窃取效果上有了很大提升,由于公开数据集先验知识的引入,促进了生成器的有效学习,使生成器能够快速满足模型功能窃取任务的需要,有效降低了查询代价。然而,本文方法对生成数据的真实性约束有一定局限,窃取效果一定程度上与真实数据集的选取有关。在下一步研究中,将尝试通过迁移学习提升生成模型的泛化能力,进而减弱对公开数据集的依赖。

参考文献 (References)

- Brown T B, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler D M, Wu J, Winter C, Hesse C, Chen M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, McCandlish S, Radford A, Sutskever I and Amodei D. 2020. Language models are few-shot learners [EB/OL]. [2020-05-28]. <https://arxiv.org/pdf/2005.14165.pdf>
- Chen P Y, Zhang H, Sharma Y, Yi J F and Hsieh C J. 2017. 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, USA; Association for Computing Machinery: 15-26 [DOI: 10.1145/3128572.3140448]
- Chen T, Kornblith S, Norouzi M and Hinton G. 2020. A simple framework for contrastive learning of visual representations//Proceedings of the 37th International Conference on Machine Learning. [s. l.]: JMLR. org: 1597-1607
- Correia-Silva J R, Berriel R F, Badue C, de Souza A F and Oliveira-Santos T. 2018. Copycat CNN: stealing knowledge by persuading confession with random non-labeled data//Proceedings of 2018 International Joint Conference on Neural Networks (IJCNN). Rio de Janeiro, Brazil: IEEE: 1-8 [DOI: 10.1109/IJCNN.2018.8489592]
- Deng J, Dong W, Socher R, Li L J, Li K and Li F F. 2009. ImageNet: a large-scale hierarchical image database//Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, USA: IEEE: 248-255 [DOI: 10.1109/CVPR.2009.5206848]
- Duchi J C, Jordan M I, Wainwright M J and Wibisono A. 2012. Finite sample convergence rates of zero-order stochastic optimization methods//Proceedings of the 25th International Conference on Neural Information Processing Systems. Lake Tahoe, USA: Curran Associates Inc.: 1439-1447
- Fang G F, Song J, Shen C C, Wang X C, Chen D and Song M L. 2020. Data-free adversarial distillation [EB/OL]. [2020-05-02]. <https://arxiv.org/pdf/1912.11006.pdf>
- Li F F, Fergus R and Perona P. 2004. Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories//Proceedings of 2004 Conference on Computer Vision and Pattern Recognition Workshop. Washington, USA: IEEE: #178 [DOI: 10.1109/CVPR.2004.383]
- Gong X L, Chen Y J, Yang W B, Mei G H and Wang Q. 2021. InverseNet: augmenting model extraction attacks with training data inversion//Proceedings of the 13th International Joint Conference on Arti-

- ficial Intelligence. Montreal, Canada; IJCAI. org; 2439-2447 [DOI: 10.24963/ijcai.2021/336]
- Goodfellow I J, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A and Bengio Y. 2014. Generative adversarial nets//Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal, Canada; MIT Press; 2672-2680
- Hadsell R, Chopra S and LeCun Y. 2006. Dimensionality reduction by learning an invariant mapping//Proceedings of 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. New York, USA; IEEE; 1735-1742 [DOI: 10.1109/CVPR.2006.100]
- He K M, Fan H Q, Wu Y X, Xie S N and Girshick R. 2020. Momentum contrast for unsupervised visual representation learning//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA; IEEE; 9726-9735 [DOI: 10.1109/CVPR42600.2020.00975]
- He K M, Zhang X Y, Ren S Q and Sun J. 2016. Deep residual learning for image recognition//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA; IEEE; 770-778 [DOI: 10.1109/CVPR.2016.90]
- He P S, Li W C, Zhang J Y, Wang H X and Jiang X H. 2022. Overview of passive forensics and anti-forensics techniques for GAN-generated image. *Journal of Image and Graphics*, 27(1): 88-110 (何沛松, 李伟创, 张婧媛, 王宏霞, 蒋兴浩. 2022. 面向GAN生成图像的被动取证及反取证技术综述. *中国图象图形学报*, 27(1): 88-110) [DOI: 10.11834/jig.210430]
- Hinton G, Vinyals O and Dean J. 2015. Distilling the knowledge in a neural network [EB/OL]. [2020-05-09]. <https://arxiv.org/pdf/1503.02531.pdf>
- Kariyappa S, Prakash A and Qureshi M K. 2021. MAZE: data-free model stealing attack using zeroth-order gradient estimation//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA; IEEE; 13809-13818 [DOI: 10.1109/CVPR46437.2021.01360]
- Karras T, Laine S and Aila T. 2019. A style-based generator architecture for generative adversarial networks//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA; IEEE; 4396-4405 [DOI: 10.1109/CVPR.2019.00453]
- Krizhevsky A. 2009. Learning Multiple Layers of Features from Tiny Images. Toronto; University of Toronto
- Li Y and Song P H. 2022. Review of transfer learning in medical image classification. *Journal of Image and Graphics*, 27(3): 672-686 (黎英, 宋佩华. 2022. 迁移学习在医学图像分类中的研究进展. *中国图象图形学报*, 27(3): 672-686) [DOI: 10.11834/jig.210814]
- Netzer Y, Wang T, Coates A, Bissacco A, Wu B and Ng A Y. 2011. Reading digits in natural images with unsupervised feature learning. *NIPS Workshop on Deep Learning & Unsupervised Feature Learning (2011)*; 1-9
- Oh S J, Schiele B and Fritz M. 2019. Towards reverse-engineering black-box neural networks//*Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Cham; Springer; 121-144 [DOI: 10.1007/978-3-030-28954-6_7]
- Oreondy T, Schiele B and Fritz M. 2019. Knockoff nets: stealing functionality of black-box models//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA; IEEE; 4949-4958 [DOI: 10.1109/CVPR.2019.00509]
- Papernot N, McDaniel P, Goodfellow I, Jha S, Celik Z B and Swami A. 2017. Practical black-box attacks against machine learning//Proceedings of 2017 ACM on Asia Conference on Computer and Communications Security. Abu Dhabi, United Arab Emirates; Association for Computing Machinery; 506-519 [DOI: 10.1145/3052973.3053009]
- Quattoni A and Torralba A. 2009. Recognizing indoor scenes//Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, USA; IEEE; 413-420 [DOI: 10.1109/CVPR.2009.5206537]
- Radford A, Metz L and Chintala S. 2016. Unsupervised representation learning with deep convolutional generative adversarial networks [EB/OL]. [2020-01-07]. <https://arxiv.org/pdf/1511.06434.pdf>
- Strubell E, Ganesh A and McCallum A. 2019. Energy and policy considerations for deep learning in NLP [EB/OL]. [2020-06-05]. <https://arxiv.org/pdf/1906.02243.pdf>
- Tramèr F, Zhang F, Juels A, Reiter M K and Ristenpart T. 2016. Stealing machine learning models via prediction APIs//The 25th USENIX Conference on Security Symposium. Austin, USA; USENIX Association; 601-618
- Truong J B, Maini P, Walls R J and Papernot N. 2021. Data-free model extraction//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA; IEEE; 4769-4778 [DOI: 10.1109/CVPR46437.2021.00474]
- van den Oord A, Li Y Z and Vinyals O. 2019. Representation learning with contrastive predictive coding [EB/OL]. [2020-06-22]. <https://arxiv.org/pdf/1807.03748.pdf>
- Wang B H and Gong N Z. 2018. Stealing hyperparameters in machine learning//2018 IEEE Symposium on Security and Privacy (SP). San Francisco, USA; IEEE; 36-52 [DOI: 10.1109/SP.2018.00038]
- Xiao H, Rasul K and Vollgraf R. 2017. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms [EB/OL]. [2020-09-15]. <https://arxiv.org/pdf/1708.07747.pdf>

作者简介



李延铭, 1997 年生, 男, 硕士研究生, 主要研究方向为机器学习。

E-mail: lym@bit.edu.cn



李长升, 通信作者, 男, 教授, 主要研究方向为机器学习。

E-mail: lcs@bit.edu.cn

余佳奇, 男, 硕士研究生, 主要研究方向为机器学习。

E-mail: yujiaqi@bit.edu.cn

袁野, 男, 教授, 主要研究方向为数据库。

E-mail: yuan-ye@bit.edu.cn

王国仁, 男, 教授, 主要研究方向为数据库。

E-mail: wanggrbit@126.com