语音识别系统对抗样本攻击及防御综述

台建玮[1,2], 李亚凯[1,2], 贾晓启[1,2], 黄庆佳[1,2]

[1] 中国科学院信息工程研究所 北京 中国 100093
[2] 中国科学院大学网络空间安全学院 北京 中国 100049

摘要 语音是人类与智能手机或智能家电等现代智能设备进行通信的一种常用而有效的方式。随着计算机和网络技术的显著进步, 语音识别系统得到了广泛的应用, 它可以将用户发出的语音指令解释为智能设备上可以理解的数字指令或信号, 实现用户与这些设备的远程交互功能。近年来, 深度学习技术的进步推动了语音识别系统发展, 使得语音识别系统的精度和可用性不断提高。然而深度学习技术自身还存在未解决的安全性问题, 例如对抗样本。对抗样本是指在模型的预测阶段, 通过对预测样本添加细微的扰动, 使模型以高置信度给出一个错误的目标类别输出。目前对于对抗样本的攻击及防御研究主要集中在计算机视觉领域而忽略了语音识别系统模型的安全问题, 当今最先进的语音识别系统由于采用深度学习技术也面临着对抗样本攻击带来的巨大安全威胁。针对语音识别系统模型同样面临对抗样本的风险, 本文对语音识别系统的对抗样本攻击和防御提供了一个系统的综述。我们概述了不同类型语音对抗样本攻击的基本原理并对目前最先进的语音对抗样本生成方法进行了全面的比较和讨论。同时, 为了构建更安全的语音识别系统, 我们讨论了现有语音对抗样本的防御策略并展望了该领域未来的研究方向。

关键词 语音识别系统; 语音对抗样本; 防御策略; 深度学习
中图法分类号 TP309.2 **DOI 号** 10.19363/J.cnki.cn10-1380/tn.2022.09.05

A Survey: Attacks and Countermeasures of Adversarial Examples for Speech Recognition System

TAI Jianwei[1,2], LI Yakai[1,2], JIA Xiaoqi[1,2], HUANG Qingjia[1,2]

[1] Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China
[2] School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100049, China

Abstract Voice is a common and effective way of communication between human and modern intelligent devices such as smartphones or smart household appliances. With the significant progress of computer and network technology, the speech recognition system has been widely used. It can interpret the voice instructions sent by users into understandable digital instructions or signals on intelligent devices and realize the remote interaction between users and intelligent devices. In recent years, the impressive achievements of deep learning technology promote the development of the speech recognition system, which makes the accuracy and availability of speech recognition system improve continuously. However, deep learning technology itself still has unsolved security problems, such as adversarial examples. Adversarial samples refer to adding subtle perturbations to the predicted samples in the prediction stage of the model, make the model gives a wrong target category classify output with high confidence. The current researches on the attack and defense of adversarial samples mainly focuses on the field of computer vision and ignoring the security issues of the speech recognition systems model. At present, the most advanced speech recognition system also faces a huge security threat brought by adversarial examples attack due to the use of deep learning technology. In response to the same risk of adversarial samples faced in the field of speech recognition system, this paper provides a systematic overview of attacks and countermeasures of adversarial examples for the speech recognition system. First of all, we summarize the basic attack principles of different types of speech adversarial examples. In addition, we discuss the advantages and disadvantages of these methods, through a comprehensive comparison of the most advanced generation methods of speech adversarial examples. Last but not least, in order to build a more secure speech recognition system, we discuss the defense countermeasures for the existing speech adversarial examples and look forward to the future research direction in this field.

Key words speech recognition system; adversarial examples; defense strategies; deep learning

通讯作者: 贾晓启, 博士, 研究员, Email: jiaxiaoqi@iie.ac.cn。
本课题得到中国科学院网络测评技术重点实验室资助项目, 网络安全防护技术北京市重点实验室资助项目, 北京市科技计划课题(No. Z191100007119010), 国家自然科学基金(No. 61772078)资助。
收稿日期: 2019-12-31; 修改日期: 2020-03-06; 定稿日期: 2022-07-14

1 引言

语音识别是一种能够使得智能设备识别和理解人类语音的技术, 由于先进的语音识别系统都具备自动化能力, 因此也称为自动语音识别[1]。得益于语音识别技术近年来在识别精度上的大幅提升, 语音识别以其易用性和高效性成为了一种越来越流行的人机交互机制。因此, 语音识别将各种各样的软件服务转变为了可语音控制的系统和智能音响也逐渐走入了普罗大众的家庭生活。除了亚马逊 Alexa、谷歌助手、苹果 Siri、讯飞听见等商业产品外, 还有 Kaldi[2]、卡内基梅隆大学的 Sphinx[3]和 Mozilla DeepSpeech[4]等开源平台。图 1 概述了一个典型的语音识别系统架构, 包括两个主要组成部分: 音频获取模块和语音模型。其中音频获取模块由音频采集设备和信号处理设备组成。语音模型有三个子模块组成: 特征提取模块, 声学模型和语言模型。原始音频经过功率放大器和滤波器后, 语音识别系统需要从数字化的音频信号中提取声学特征。常用的声学特征提取算法有 Mel 频率倒谱系数(mel-frequency cepstral coefficients, MFCC)[5]、线性预测系数[6]等, 其中在商业产品和开源平台中使用频率最高的均为 MFCC。同时语音信号是一种典型的时序信号, 其所含信息在时间跨度上有较大差异, 因此语音识别系统需要使用短时分析定期对语音信号进行评估。

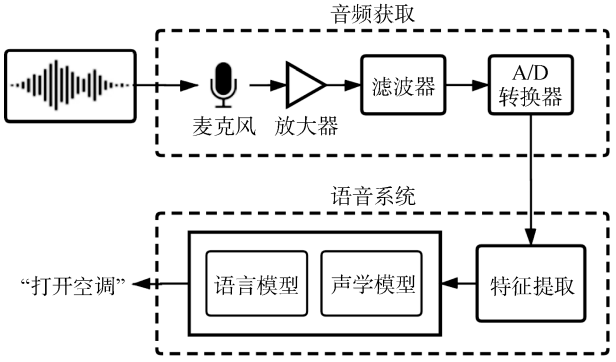

图 1 语音识别系统架构图

Figure 1 The architecture of speech recognition system

近年来, 随着深度学习技术[7]的蓬勃发展, 深度学习在各行各业中的应用不断加深。深度学习归属于机器学习大类中的一个子集, 它通过训练一个神经网络模型来执行原本需要人类理解能力的任务, 例如语音识别、图像识别或多样化样本生成等。术语“深度”通常指的是神经网络中隐藏层的数量, 原始的神经网络一般只包含 2 到 3 个隐藏层, 而深层神经网络可以包含多达上百个隐藏层。大多数深度学习方法通过建立基于数据的神经网络结构和目标函数, 利用大规模标记数据的优势, 使神经网络模型直接从数据中学习特征, 而不需要手动提取特征, 达到神经网络模型自动化学习的目的。因此该类方法可以达到比传统机器学习方法更高的识别准确率。最近的研究表明, 深度学习在一些特定任务中甚至比人类做的更好, 例如图像识别。因此, 通过将深度学习技术引入语音识别这一任务领域, 研究人员实现了语音识别系统识别精度的大幅提升, 进一步推动了语音识别系统的大规模应用。

尽管深度学习技术对语音识别任务的精度提升具有重要价值, 但深度学习技术也带来了巨大的安全问题, 其中对深度学习模型的安全威胁最大的就是对抗样本[8]攻击, 这一点已经在图像领域得到了广泛的研究和证明。对抗样本是由攻击者精心设计的具有误导能力的恶意样本。例如, 攻击者在一张熊猫图片中引入一个微小的扰动后, 人类对该扰动毫无感知, 但是基于深度学习的图像识别模型确以 99.3%的置信度将熊猫图片识别为长臂猿[9]。这种对抗样本攻击给基于深度学习的应用带来了巨大的安全风险。 例如, 攻击者可以针对自动驾驶汽车使用贴纸或油漆创建一个具有对抗性的停车标志, 而车辆基于图像识别的自动驾驶系统将该停车标志识别为“加速”或其他标志, 造成极大的危害。近年来, 针对深度学习中对抗样本的研究越来越受到研究人员的关注。但是大多数研究集中在图像领域, 对语音识别系统乃至整个语音领域的安全性研究还存在不足。

随着深度学习模型在语音识别任务上的精度不断提高, 目前最先进的语音识别系统大都采用深度学习作为核心技术, 因此攻击该系统的关键点在于神经网络容易受到对抗样本干扰和欺骗这一安全漏洞[10]。这一点自然而然地激发了攻击者构建语音对抗样本的灵感。攻击者只需要在原语音上添加微小的扰动, 使得人耳察觉不到扰动的存在, 但语音识别系统的预测结果却可以受到攻击者的干扰和误导。但是在实际的攻击场景中, 对抗扰动的添加往往会造成明显的噪声或在空气信道传播时被环境噪声破坏。因此, 有效地生成对抗扰动变得更具有挑战性。因此, 生成语音对抗样本的过程往往需要在扰动复杂性, 攻击有效性和攻击隐蔽性之间进行权衡。由于语音对抗样本的巨大研究潜力, 针对语音对抗样本攻防两端的研究工作不断涌现, 已经成为了一个新兴的研究热点。探索针对语音对抗样本的防御策

略, 实现更安全的语音识别系统是研究语音对抗样本攻击方法的主要目标。根据相关工作的研究进展, 本文主要分析和讨论对语音识别系统中声学模型的攻击和防御问题。

2 关键技术问题

2.1 语音对抗样本

近年来, 深度学习在图像分类、语音识别、自动驾驶等领域中得到了广泛的应用, 虽然它显著地提高了目标任务的精度, 但也不可避免的面临着对抗样本带来的安全威胁。由于深度学习在语音识别任务中极好的适用性, 目前最先进的语音识别系统大都基于深度学习技术来实现其声学模型和语音模型。因此, 目前语音识别系统面临的最大安全威胁就是语音对抗样本攻击。

从概念上说, 一个深度学习模型本质上是一个映射函数, 它将模型输入映射到相应的概率输出, 此时找到一个与模型输入相近但其概率输出与该输入差异较大的样本, 此时这两个输入的差异太小以至于人类无法对其区分, 但其模型输出概率能够将它们识别为不同的类别。为了欺骗语音识别系统, 攻击者通过在正常语音样本上添加精心设计的微小扰动来生成语音对抗样本。人类无法听见这种扰动或该扰动仅仅被认为是微弱的背景噪声, 但却使得语音识别模型预测错误, 甚至误导该模型产生攻击者预期的结果, 从而导致目标系统接受恶意的控制指令。图 2 展示了语音对抗样本[11]对目标语音识别系统的攻击过程。针对对抗样本的原理, Biggio 等人[12]描述了相关的几个重要概念, 包括对抗目标(Adversary's goal)、对抗知识(Adversary's knowledge)和对抗能力(Adversary's capability)等。这些概念同样适用于语音对抗样本。

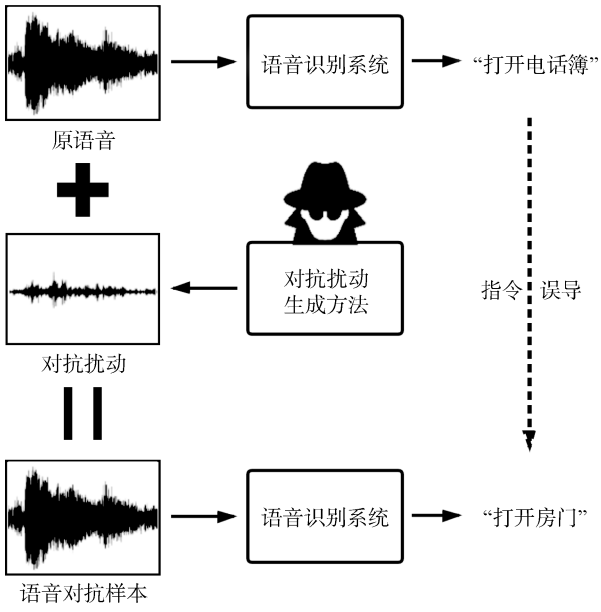

图 2 语音对抗样本攻击流程

Figure 2 The attack pipeline of speech adversarial examples

2.2 语音对抗样本威胁模型

根据语音对抗样本的攻击在知识, 目标和背景上的差异, 我们从以下多个维度来讨论该攻击威胁模型: 对抗知识, 对抗目标性, 攻击对象和语音对抗载体。

2.2.1 对抗知识

根据攻击者在进行对抗攻击时对目标系统的知识掌握程度的不同, 可以将语音对抗攻击分为白盒攻击和黑盒攻击两种类型。白盒攻击假设攻击者对目标深度学习模型有较为全面的了解, 包括模型类型, 模型架构以及所有预训练参数和权重等, 同时攻击者可以不受限制地与目标模型进行交互。而黑盒攻击则假设攻击者只掌握了很少的目标模型知识, 甚至是只能得到模型输出结果的普通用户。因此, 黑盒攻击的难度和实用性都远高于白盒攻击。

2.2.2 对抗目标性

根据攻击者在对语音识别系统进行语音对抗样本攻击时是否有预期的目标指令, 可以将语音对抗样本攻击分为无目标攻击和有目标攻击两种类型。无目标攻击旨在使深度学习模型为语音对抗样本预测任何不正确的分类, 即该攻击生成的语音对抗样本只能干扰目标语音识别系统的正常功能而不能完成其它特定的攻击意图。而有目标攻击是一种更强大的攻击方式, 它旨在误导深度学习模型为语音对抗样本预测特定的分类, 该特定的分类由攻击者设定。 这种攻击不只可以干扰语音识别系统的正常功能, 而且可以使得被攻击系统执行符合攻击者意图的恶意指令, 引发严重的安全问题。

2.2.3 攻击对象

攻击对象是指语音对抗攻击的目标系统。在针对语音识别系统的语音对抗样本攻击中, 其攻击的对象主要是各类白盒或黑盒的语音识别系统。其中白盒系统除了为语音识别任务定制化的模型外, 还有很多开源的白盒系统, 比较常见的有: Mozilla DeepSpeech[4], DeepSpeech-2[13]和 Kaldi[2]等。而黑盒语音识别系统主要包含商用语音 API 或语音物理设备, 其中商用语音 API 在各大科技厂商均有相关产品, 而语音物理设备较为常见的有 Google Home, Amazon Echo 和 Microsoft Cortana 等。需要指出的是, 本文将具有语音识别系统的物理设备看做黑盒系统,

并以此作为目标系统来讨论。

2.2.4 语音对抗载体

对于语音识别系统的对抗样本来说, 其对抗样本载体, 即对抗样本的存在形式极大的影响了语音对抗样本的隐蔽性。一般来说, 语音对抗样本载体主要为静默, 噪声和错义语音这几种音频类型[14]。静默的语音对抗样本由于其语音频段超过人耳的感知范围, 因此对用户具有强大的隐蔽性。而噪声相对于静默来说, 更容易被人耳感知, 因此隐蔽性较差, 但生成这类语音对抗样本的效率高, 难度低。错义语音相对于前两种类型来说更为常用, 它利用心理声学的知识降低了人耳对语音对抗样本的敏感性, 即人耳在接收一个明确语音信息的同时会忽略其中的扰动。攻击者通过微调原语音生成该类语音对抗样本, 并保证用户无法察觉该样本与原语音的差异, 同时可以误导语音识别系统进行错误分类或实现其他特定攻击意图。

2.3 语音对抗样本可迁移性

有趣的是, 在训练集的不同子集上训练的具有不同权重参数的模型都会对相同的对抗样本产生错误的分类结果, 即对抗样本具有可迁移性。这也说明了对抗样本不单纯是模型欠拟合的问题, 而是深度学习技术本身的一个盲点。

Wei 等人[15]验证了对抗样本的可迁移性, 即针对一种深度学习模型生成的对抗样本在迁移到另一个不同结构的模型后仍然具备一定的攻击能力。由于目前先进的语音识别系统大多基于深度学习模型, 因此语音对抗样本也继承了可迁移性这一特点, 这为语音对抗样本迁移攻击提供了理论基础。为了定量化地评估不同对抗样本的可迁移性, 他们提出一种迁移成功率的计算方式。即给定两个模型, 通过计算一个模型生成的对抗样本在另一个模型上分类正确的百分比, 用来评估无目标攻击的可转移性。更低的比率说明对抗样本的可迁移性更好。而评估有目标攻击的可迁移性通过计算一个模型生成的对抗样本在另一个模型中分类为攻击者预期类别的比率, 该比率越高代表对抗样本的可迁移性越好。通过以上两种方法, 可以对对抗样本的可迁移性进行较为全面的评估。

2.4 攻击方法分类

根据攻击方式的差异, 可以从不同的角度对语音对抗样本的攻击方法进行归类。本文从语音对抗样本攻击信道的角度, 根据其攻击信道差异, 即语音对抗样本通过数字信道或物理(空气)信道的差异, 将语音对抗样本攻击方法分为两个大类: 数字攻击和物理攻击。

数字攻击是指生成的语音对抗样本直接通过数字的方式输入目标语音识别系统, 这种攻击不需要通过空气信道传输语音对抗样本, 因此不需要考虑环境噪声对对抗扰动带来的影响。一般情况下, 数字攻击可以根据对抗扰动生成算法的不同具体地归纳为 4 种主要类型: 基于梯度符号, 基于迭代优化, 基于遗传算法和通用化扰动。

物理攻击是指生成的语音对抗样本需要通过空气信道传播, 再由语音采集设备(例如麦克风)将语音信号转化为系统可理解的数字信号, 之后再输入语音识别系统。由于这种攻击要求语音对抗样本通过空气信道传播, 因此对样本的鲁棒性提出了更高的要求。空气信道中包含了大量未知的噪声信号, 这会对样本中的对抗扰动产生很大的干扰, 导致语音对抗样本失去攻击能力。为此, 物理攻击方法需要考虑空气信道这一影响因素。同时, 物理攻击方法需要关注语音对抗样本对于人耳的隐蔽性, 需要将对抗扰动控制在人耳不可感知的频率范围。更具体的, 物理攻击根据目标系统的特征还可以具体分为语音接口攻击和语音设备攻击两种类型。

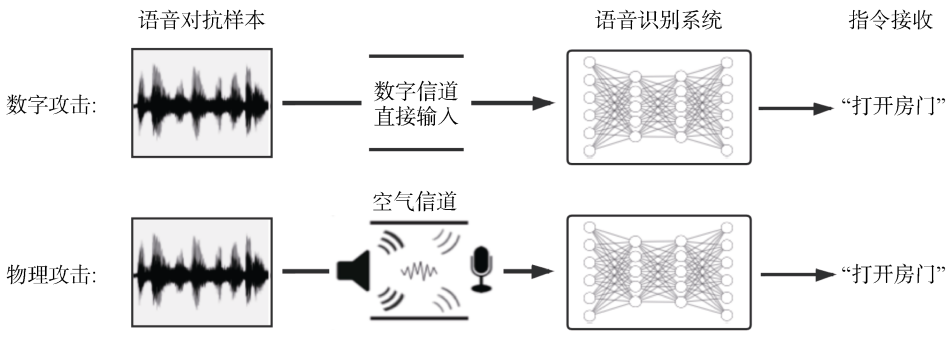

图 3 数字攻击与物理攻击

Figure 3 Digital attack and physical attack

2.5 防御策略分类

攻击方法与防御策略的关系就像是硬币的两面, 两者互相竞争又互相促进。与对抗攻击分类相比, 语音对抗样本的防御策略更为丰富。目前, 根据思路和切入点的差异可以将语音对抗样本的防御策略划分为以下几个主要类别: 对抗训练, 对抗样本检测, 数据压缩防御和模型优化。对抗样本检测根据特征来检测并过滤出语音对抗样本, 以此语音识别系统的防御能力。值得指出的是, 对抗训练(Adversarial Training)一般是指使用对抗性目标函数, 即最大化对抗分类器目标, 来优化模型输出的概念。区别于上述概念, 此处对抗训练的概念是指通过将语音对抗样本加入模型的训练过程(训练集), 来提升语音识别模型的鲁棒性, 进而提升对语音对抗样本的防御能力。数据压缩防御从破坏对抗扰动的攻击能力的角度缓解语音对抗样本带来的安全威胁。而模型优化则是通过额外的辅助模块来改进模型自身结构, 降低语音识别系统的脆弱性。

3 语音对抗样本攻击方法

由于针对语音识别系统的攻击和防御是对立统一的, 因此对语音识别系统的各类攻击方法进行全面和系统的研究是提升语音识别系统安全性的必要步骤。针对语音识别系统的攻击有很多种, 目前根据攻击信道的不同可以归纳为两种主要类别, 即数字攻击和物理攻击, 如图 4 所示。在这一节中, 对于不同类别中的每一个相关工作, 本文将简要回顾其攻击方法并对这些方法进行全面的比较和讨论。

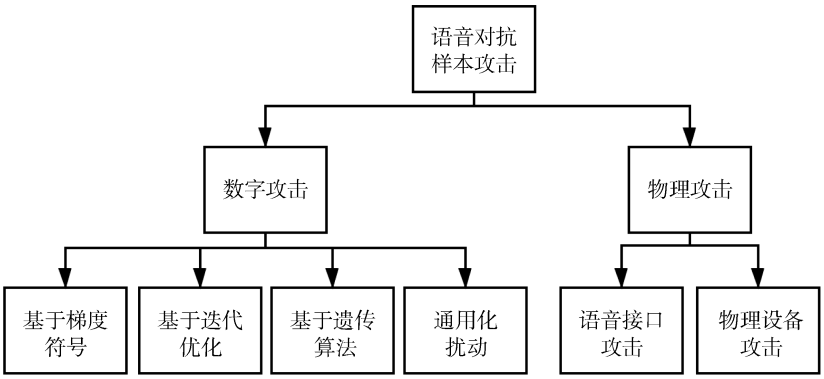

图 4 语音对抗样本攻击分类

Figure 4 The attack categories of speech adversarial examples

3.1 数字攻击

3.1.1 基于梯度符号

Gong 等人[16]提出了一种基于梯度符号的语音识别系统对抗样本生成方法, 适用于白盒攻击场景。该方法通过直接扰动音频记录的原始波形而不是特定的声学特征来生成用于误导语音识别系统的语音对抗样本。实验表明该方法生成的对抗扰动可以导致最先进的语音识别系统性能显著下降, 而对原语音的质量影响较小。为了避免声学特征转换回波形而带来的感知损失, 他们提出了一种针对端到端的深度学习模型的扰动方法来直接修改原始波形并且使用卷积层代替递归结构来解决梯度消失问题。在干扰因子为 0.032 的情况下, 攻击成功率提高了 30%左右。然而语音对抗样本的有效性在很大程度上依赖于替代网络的训练, 因此通过对抗样本的可转移性进行黑箱攻击的可能性很小, 实用价值有限。

同样的, Kreuk 等人[17]将基于梯度符号的方法应用于声学特征(例如 MFCC), 然后根据该声学特征重建音频波形。通过将假阳性率提高大约 90%, 可以大大提高攻击性能, 这显然比 Gong 等人[16]的方案更好。此外, 该攻击进行了两次黑盒攻击验证, 证明所该方法生成的对抗扰动具有可迁移性, 但是没有对对抗扰动和攻击准确性进行精确评估。同时, 目前基于梯度符号方法的语音对抗样本生成技术的研究相对较少。

3.1.2 基于迭代优化

相比于基于梯度符号的方法, 基于迭代优化的方法也是利用了梯度来计算所需的对抗扰动。但是基于迭代优化的方法使用优化器不断执行迭代过程, 可以实现更加快速和细粒度的对抗扰动计算。一般情况下, 基于迭代优化方法要求攻击者对目标模型的信息有充分的了解, 因此攻击前提较为苛刻, 但是通常比其他方法拥有更高的攻击成功率。

传统的语音识别系统由不同的组件组成, 例如特征提取模块, 声学模型和语言模型等, 其中每个组件都是单独设计和训练的。最近, 语音识别系统研

究集中于基于深度学习的端到端模型设计, 这类模型无需输入额外的预处理信息即可获得语音特征并输出识别结果, 提升识别任务的效率。Cisse 等人[18]介绍了 Houdini 攻击, 一种针对黑盒系统的对抗样本生成方法, 该攻击通过生成能够直接导致目标系统丧失识别能力的语音对抗样本来攻击任何基于梯度的语音识别模型。在实验中, 该攻击通过生成人类无法与察觉的语音对抗样本(经 ABX 实验验证), 对 DeepSpeech-2[13]深度语音识别模型进行了成功的无目标攻击。同时, Cisse 等人[18]还通过在黑盒攻击场景中对 Google Voice[19]语音识别系统进行攻击来研究语音对抗样本的可迁移性。

与上述两种方法类似, Carlini 等人[20]也构建了一种针对语音识别模型的对抗样本攻击。对于给定的任何音频, 该攻击只需添加小于 0.01%的对抗扰动就可以使得目标语音识别系统将其理解为任何预期的指令。该攻击利用基于迭代优化的方法, 针对端到端的白盒系统, 直接对用作语音识别模型输入的原始样本进行操作来添加对抗扰动。但是这需要通过困难的 MFCC 逆变换来实现, 因此该攻击将 MFCC 预处理过程重新实现以保证该过程可求梯度, 实现对原语音进行直接修改。总的来说, 该攻击基于白盒的迭代优化方法生成对抗样本, 并在先进的语音识别模型 DeepSpeech[4]中达到了 100%的攻击成功率。

3.1.3 基于遗传算法

假设攻击者知道模型的架构和参数, 可以使用反向传播有效地计算出对抗扰动所需的精确梯度。但是基于链式法则的梯度计算需要具备计算模型所有网络层梯度的能力, 尽管这一要求在图像识别模型中容易做到, 但是将相同的技术应用于语音识别模型却变得困难, 因为它们大多依赖于信号处理模块的输出(如频谱和 MFCC)作为输入语音数据的特征。而提取频谱和 MFCC 的模块往往是不可微的, 没有有效的方法来计算它们的梯度, 此时对抗性扰动的计算面临困境。因此, 研究人员另辟蹊径, 尝试通过引入遗传算法来避免计算特征提取模块的梯度。遗传算法是一种通过模拟自然进化过程搜索最优解的启发式优化算法[21], 其主要特点是直接对结构对象进行操作, 不存在求导和函数连续性的限定。该算法在创建一系列具有代表性的对抗样本之后, 适应性更高的候选样本更有可能变异并成为下一代的一部分, 重复迭代过程并得到最终结果。

基于遗传算法不依赖梯度优化的这一优点, Alzantot 等人[22]提出了第一个基于遗传算法的解决方案来生成语音对抗样本。该攻击从创建大量候选样本开始, 然后为每个候选样本计算适合度分数, 得分更高的样本更容易变异。此方案以 87%的成功率实现了对语音识别系统有目标的黑盒攻击, 并且 89%的人类测试者将这些对抗扰动视为语音背景噪声。但是该方法仅在单个单词的语音片段上进行了性能评估, 而对于语句级别的语音片段是否有效尚不明确。更重要的是, 该攻击仅在定制的深度学习模型上有效, 面对其他高级语音识别系统时无法保证攻击成功率。

受到上述研究工作的启发, Taori 等人[23]通过结合遗传算法和梯度估计提出了一种性能更加优越的, 针对黑盒语音识别模型的语音对抗样本生成方法。该方法可以在对抗扰动陷入局部最小值时加快收敛速度并增加突变概率。为了限制过多的突变, 从而限制过多的噪声, Taori 等人[23]设计了新的动量突变更新算法来改进标准的遗传算法。攻击的第二阶段使用梯度估计来计算各个语音点的梯度, 因此在语音对抗样本接近预期目标时, 可以更准确地添加扰动。该攻击采用黑盒方法进行语音对抗样本生成, 实现对语音识别系统有目标的攻击以执行预期的恶意指令。在遗传迭代进行了 3000 次之后, 与预期目标指令达到了 89.25%的相似性, 同时与原语音保持了 94.6%的相似性。然而, 对流行的 DeepSpeech[4]模型进行攻击时, 该方法的攻击成功率只有 35%。

3.1.4 通用化扰动

与此同时, 语音领域也存在具有通用性的对抗扰动。Neekhara 等人[24]验证了通用化语音对抗扰动的存在, 这些扰动会通过语音识别系统引起语音信号的错误识别。由于图像领域的通用性对抗扰动算法不能在语音领域直接使用, 因此他们解决了一个替代的优化问题以实现针对语音识别系统的通用性对抗性扰动生成方法。为了将每个数据点推向其决策边界, 即将大多数数据点推到语音识别模型的正确分类区域之外, 该算法通过迭代遍历训练数据集以构建具有通用能力的扰动向量。最终 Neekhara 等人[24]提出了一种白盒攻击算法来寻找单个不可感知的语音对抗扰动, 将其添加到任意语音信号中以生成语音对抗样本, 并使用通用化扰动成功对先进语音识别模型 DeepSpeech[4]发起了攻击。同时, 通过在基于 WaveNet 的语音识别系统上进行语音对抗样本迁移攻击测试, 该通用性语音对抗扰动生成方法被证明在很大程度上可在具有不同架构的模型之间迁移。

Abdoli 等人[25]也提出了一种的通用性对抗扰动的生成方法, 这种通用性扰动可以欺骗针对有目标

和无目标攻击的语音识别系统。他们提出了两种实现这种通用对抗扰动的方法。第一种方法基于图像领域众所周知的迭代贪婪算法: 它将细微扰动聚集到输入以便将其推到决策边界。第二种方法是这项工作的主要技术贡献, 即一种新颖的惩罚公式。与贪婪算法不同, 惩罚方法使一批样本上的目标函数最小化。因此, 当训练样本的数量有限时, 它将产生具有更高成功率的攻击。实验证明该方法的有目标攻击和无目标攻击的攻击成功率分别高于 91.1%和74.7%。值得注意的是, 无论是有目标攻击还是无目标攻击, 在训练集上生成的对抗扰动可以很好地迁移到测试集上。

3.2 物理攻击

在语音直接以数字的方式输入目标语音识别系统的情况下, 攻击者通过使用能够清晰描述对抗目标的攻击算法来确定目标系统的数据点位置来生成语音对抗样本。但在语音通过空气信道输入目标系统的情况下, 语音对抗样本攻击的难度将会大大增加, 这种困难可以归因于环境的混响和来自扬声器和麦克风的噪音。相比之下, 通过空气信道传播的语音对抗样本需要对未知的环境和设备具有更好的鲁棒性。同时, 考虑到空气信道的独特性, 语音对抗样本可以同时干扰大量的目标系统, 实现大规模攻击。因此, 使用该信道传播语音对抗样本的安全威胁将远远大于直接以数字的方式输入目标语音识别系统的场景。更具语音识别服务提供方式的不同, 我们还可以将物理攻击分为语音接口攻击和物理设备攻击。语音接口攻击是指目标语音系统是以白盒模型或黑盒 API 的形式提供识别服务, 攻击者可以通过数字接口与目标系统进行直接的交互。而物理设备假设攻击对象是真实物理世界中的商用语音设备, 这些设备只通过空气信道与用户交互。此时, 目标系统对于攻击者来说是完全的黑盒, 无法了解模型信息, 甚至无法通过接口与识别服务进行交互。

3.2.1 语音接口攻击

Yuan 等人[26]提出通过基于迭代优化的方法生成有效的对抗扰动并将该扰动添加到音乐中实现对语音识别系统的攻击。与 Carlini 等人[21]类似, Yuan 等人也使用可逆的 MFCC 提取模块实现对原语音信号波形的修改。该攻击考虑到播放音乐的普遍性以及将语音对抗样本隐藏在多样化的音乐波形后面的便利性, 因此选择了不同类型的音乐作为对抗攻击的载体。 通过对歌曲和预期语音命令在声学模型中的输出特征, 不断执行梯度下降来生成具有最小化扰动的语音对抗样本以保证对用户的隐蔽性。在 Kaldi[2] 上的攻击实验表明, 该攻击能够达到100%的成功率, 并且对原音乐的信噪比干扰很小。此外, 他们首次展示了语音对抗样本通过空气信道执行攻击的可能性。具体来说, 为了保证这种通过空气信道的语音对抗样本切实可行, 该攻击集成了一个通过用户说话时产生的电子噪声得到的通用噪声模型。因此, 这种具有对抗攻击能力的音乐可以通过空气信道进行传播而不丢失攻击者预期的恶意指令。但是该攻击实验是在短距离(即 1.5 米)内进行的, 因此需要进一步探索在真实攻击场景中的攻击性能。

无独有偶, Yakura 等人[27]也提出一种方法来产生一种具有良好鲁棒性的语音对抗样本, 可以在真实世界中攻击语音识别模型。该方法通过模拟物理世界中的回放或录制所引起的信号变换, 然后将这些变换进行建模并添加到语音对抗样本的生成过程中, 使得语音对抗样本对空气信道的鲁棒性更好。该方法通过空气信道播放的语音对抗样本对 DeepSpeech[4]发起了成功的攻击, 并且保证该攻击不能被用户察觉。

上述工作已经实现了更具鲁棒性的语音对抗样本并在真实世界中成功发起了攻击, 但是语音对抗样本对用户的隐蔽性还存在不足。尽管目前生成语音对抗样本的工作在实现方法上都考虑了对对抗扰动的最小化, 但是人耳相比于人眼更加敏感, 语音中的微小扰动也容易被用户发现。因此, 为了缓解人耳对语音对抗样本对的敏感性, Qin 等人[28]利用人耳掩蔽这一心理声学原理, 实现了一种新颖的生成有效的且不可察觉的语音对抗样本的方法。具体来说, 该方法利用人耳掩蔽的心理声学原理, 只在人类感知不到的频率区域添加对抗性扰动。例如在一个高频信号之后添加相对频率更低的对抗扰动, 即使该扰动在绝对能量方面不是“安静的”, 但仍然能够保证对人耳“隐形”。该方法通过人类听觉实验验证了对人耳的隐蔽性, 同时生成了具有完整句子的语音对抗样本, 并在 Lingvo[29]语音识别系统上实现了100%的攻击成功率。

Szurley 等人[30]也注意到人类听觉系统的心理声学特性可以被利用来生成有效但更不易察觉的对抗扰动。因此, 一种基于心理声学特性的损失函数和房间脉冲响应建模的自动化语音对抗扰动方法被提出以创建在多个物理环境(例如多个不同的房间)中可以通过空气信道传播的语音对抗样本。另外, 大多数现有语音对抗样本研究依赖于主观的人类听力测试来评估样本的质量, 这些测试没有明确解释对抗性扰动的感知失真。为了弥补这一研究空白, Szurley 等

人[30]提出一种新的评估指标, 即语音质量感知评估分数来评估语音对抗样本的质量, 这使得对攻击样本的评估不再依赖主观的人类听力测试。

尽管通过空气信道传播的语音对抗样本已经被证明能够达到很高的攻击成功率, 但是攻击者需要有关攻击发生的物理环境的精确信息, 以便根据特定环境声学模型设置和调整对抗扰动, 因此不具备转移到其他物理环境的能力。其他通过空气信道实现强大攻击效果的语音对抗样本主要是攻击者手工制作的样本, 因此不具备实现大规模, 自动化攻击的能力。Schönherr 等人[31]提出了一种生成语音对抗样本的通用化方法, 使得语音对抗样本在空气信道中保持了鲁棒性, 通过语音转录或重放的方式对目标系统实现攻击。所提出的方法只需要粗略的估测房间的信息(长、宽、高), 使用房间脉冲响应模拟器来强化对抗扰动, 而不需要实际进入房间进行环境声采集。他们使用开源语音识别系统 Kaldi[2]演示该攻击, 并使用房间脉冲响应模拟器来强化语音对抗样本, 以应对不同的房间特征。此外, 该算法也利用心理声学知识将对原语音信号的大部分扰动隐藏在人类听觉的掩蔽区域内。通过这种方法生成的语音对抗样本具有较好的迁移能力, 能够适应不同的房间特征, 同时也可以根据特定的精确房间信息对对抗扰动进行调整。因此, 攻击者可以针对任意房间设置优化对抗扰动, 成功实现对目标系统任何类型的有目标攻击。

3.2.2 物理设备攻击

在真实的攻击场景下, 攻击者往往不是针对语音识别接口进行攻击, 而是与语音识别的物理设备(例如 Google Home 和 Amazon Echo 等)进行直接交互来实现对这些设备的语音对抗样本攻击。与语音识别接口不同, 攻击者无法获得商业的语音识别设备所使用的系统信息, 语音样本也无法通过数字的方式输入系统。所以并不清楚在这种真实物理世界中的黑盒攻击场景下, 语音对抗样本是否依然能够通过空气信道成功攻击语音识别物理设备。

针对这种真实物理世界中的黑盒挑战, Chen 等人[32]提出了一种对商业语音识别设备的进行语音对抗样本攻击的方法。作为一种新思路, 该方法的核心思想是在本地同时使用一个与目标黑盒系统大致类似的模型和一个与目标黑盒系统无关的先进白盒系统作为替代模型进行语音对抗样本的生成。这两种模型在估计目标黑盒系统的行为时可以有效地互补, 从而生成对目标黑盒系统具有高度可迁移性的语音对抗样本。更具体地说, 该方法首先使用文本到语音系统合成指令语音片段, 然后根据一定的策略将语音片段输入目标黑盒系统并更具系统输出建立一个与之类似的替代模型。这允许替代模型能够专注于对攻击最重要的数据类型, 并使得替代模型更接近目标黑盒系统。最后, 该方法将在语音数据集上训练的替代模型与开源的先进语音识别模型进行集成, 通过基于迭代优化的算法并系统地选择这两种模型交叉生成的语音对抗样本。对于 98%的攻击目标设备, 包括 Google Assistant、Google Home、Amazon Echo 和 Microsoft Cortana, 该方法至少可以生成一个符合攻击者预期恶意指令的语音对抗样本, 具有很高的实用性。这种类型的攻击方法将是未来的重要研究方向。

3.3 语音对抗攻击方法比较

一般来说, 数字攻击相较于物理攻击更容易实现, 但是攻击前提要求苛刻, 不易满足。而物理攻击则是在更真实的物理世界执行的攻击, 因此对攻击前提要求低, 实用性很强。但是这种方法往往难以实现, 甚至需要攻击者手动调整语音对抗样本。因此攻击成本很高, 无法执行大规模的攻击。

在数字攻击中, 基于遗传算法的方法是一种无梯度的优化攻击, 不依赖目标模型的信息。因此, 该方法非常适合黑盒攻击场景。然而这种方法通常需要与模型进行不断的交互以实现上千次的遗传迭代过程, 当目标语音识别模型运行在具有有限通信带宽资源的物理设备上时, 该方法耗时巨大, 甚至是不实际的。另一方面, 基于梯度符号和迭代优化的解决方案是大多属于白盒攻击, 这种攻击更为简单, 但实用性相对有限。基于梯度符号的方法是一种简化过的优化方案, 它只需一次迭代过程就能够生成对抗扰动, 比其他方法速度快很多, 但扰动的攻击效果有限。基于迭代优化的方法虽然耗时更长, 但是往往能够达到更高的攻击成功率。

在物理攻击中, 基本都是基迭代优化的方法实现对抗扰动的生成, 也都是有目标的攻击, 但是攻击对象差异较大。相比于针对语音接口的对抗攻击, 针对语音物理设备的攻击具有更高的实用价值。表 1 展示了现有语音对抗攻击方法的全面比较。本文列出了它们的目标系统, 攻击类型, 对抗扰动生成方法, 对抗知识, 对抗目标性等信息。

4 语音对抗样本防御策略

对抗样本是深度学习模型普遍存在的安全问题。对于基于深度学习的语音识别系统来说, 语音对抗样本可以在用户不可感知的前提下对系统发起攻

表 1 语音对抗攻击方法的全面比较

Table 1 The comprehensive comparison of speech adversarial examples attacks

相关工作	攻击对象	攻击类型	对抗扰动生成方法	对抗知识	对抗目标性
Gong 等人[16]	Wave CNN model	数字攻击	基于梯度符号	白盒	无目标
Kreuk 等人[17]	Customized RNN model	数字攻击	基于梯度符号	白盒&黑盒	有目标
Cisse 等人[18]	Mozilla DeepSpeech-2	数字攻击	基于迭代优化	白盒&黑盒	无目标
Carlini 等人[20]	Mozilla DeepSpeech	数字攻击	基于迭代优化	白盒	有目标
Alzantot 等人[22]	Customized CNN model	数字攻击	基于遗传算法	黑盒	有目标
Taori 等人[23]	Mozilla DeepSpeech	数字攻击	基于遗传算法	黑盒	有目标
Neekhara 等人[24]	Mozilla DeepSpeech , WaveNet	数字攻击	通用化扰动	白盒	有目标
Abdoli 等人[25]	Customized 1D-CNN Gamma, SincNet	数字攻击	通用化扰动	白盒	无目标& 有目标
Yuan 等人[26]	Kaldi	物理攻击	基于迭代优化	白盒&黑盒	有目标
Yakura 等人[27]	Mozilla DeepSpeech	物理攻击	基于迭代优化	白盒	有目标
Qin 等人[28]	Lingvo	物理攻击	基于迭代优化	白盒	有目标
Szurley 等人[30]	Mozilla DeepSpeech	物理攻击	基于迭代优化	白盒	有目标
Schönherr 等人[31]	DNN-HMM- based ASR	物理攻击	基于迭代优化	白盒	有目标
Chen 等人[32]	ASR in Google Assistant、Google Home、Amazon Echo and Microsoft Cortana	物理攻击	基于迭代优化	白盒&黑盒	有目标

击, 误导目标系统接收和执行攻击者预期的恶意指令。因此, 语音对抗样本攻击为先进的语音识别系统带来了前所未有的安全风险, 提出全面而有效的语音对抗样本防御技术是解决这一安全风险的核心问题。根据现有的研究工作, 语音对抗样本的防御策略主要包括以下四个类别: 对抗样本检测, 对抗训练, 数据压缩防御和模型优化等, 如图 5 所示。

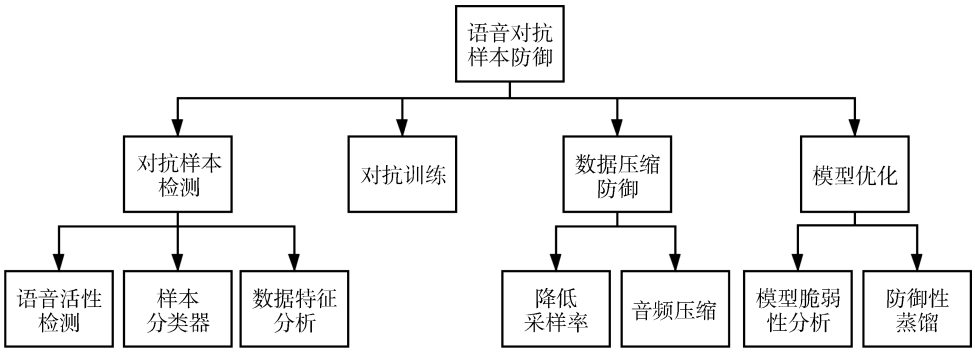

图 5 语音对抗样本防御分类

Figure 5 The countermeasure categories of speech adversarial examples

4.1 对抗样本检测

4.1.1 语音活性检测

语音活动检测是一种常用的语音处理算法, 用于检测音频样本中人类语音的存在, 在语音通讯和语音识别领域应用广泛。语音活动检测可以在输入音频信号中区分人类语音区域, 静音区域和噪声区域。 通过识别输入音频中的静音和噪声区域, 可以从该音频中消除属于这些区域的信号, 但是保持人类语音区域的有效信息。这包括消除单词之间的静音和噪声区域, 从而将整个音频转化为仅构成完整语音语义的单个单词。这对于语音识别系统尤其有用, 因为在预处理过程中使用语音活动检测可以只给模型提供必要的语音数据, 从而有可能提升系统效率并改善语音识别精度。

基于以上描述, Abdullah 等人[33]提出了一种基于语音活性检测的语音识别系统防御方案。如果在语音对抗样本攻击时先对语音的活性进行检测, 语音对抗样本可能被分类为静音或噪声, 并且无法输入目标语音模型进入下一步处理。为了验证该防御方案的有效性, 他们实现了语音活性检测算法并观察

哪些区域被识别为语音区域, 哪些被识别为静音和噪声区域。实验证明, 在各种不同情况下, 该算法均可以准确定位音频信号中的语音区域。通过语音活性检测虽然不能对语音对抗样本攻击进行完善的防御, 但它确实增加了音频无法被正确转录的可能性。即被认为是语音的噪声区域, 尤其是单词之间的噪声, 将与实际语音一起发送到语音识别系统, 这增加了部分或全部语音可能被错误翻译的概率, 从而一定程度上防御了攻击。

在语音识别系统的真实使用场景下, 来自现场用户发出的语音指令属于正常输入, 而通过其他扬声器重播的语音大概率不属于正常输入。考虑到人声的产生过程, 首先将空气从肺中排出并形成气流, 然后气流穿过气管和声带, 最后从口腔中发出而形成声波。当产生的声波和气流到达麦克风时, 不仅会捕获到语音信息, 还会捕获到一种可被感知的爆炸声, 称为爆破音。但是, 通过扬声器播放语音对抗样本不能产生由现场用户的呼吸引起的爆破音。

利用这一发声特性, Zhou 等人[34]提出可以利用用户在靠近麦克风讲话时会因为呼吸而产生一定的爆音这一现象来区分当前语音指令是确实来自用户发声还是扬声器的语音重播。通过识别语音的来源, 现场用户发出的合法指令能够被识别并输入语音识别系统进行处理, 而通过扬声器播放的语音重播将会被识别为不合法输入并在输入系统前被过滤出来, 从而实现对语音识别系统的防御方案。该方案的防御过程可以分为三个阶段: 预处理, 爆破音定位和攻击检测。预处理步骤在语音水平上将语音信号分割为音素, 以提高定位爆破音时的准确性。同时, 对所有潜在的爆破声进行了音素校正和持续时间检查。由于每个音素在人的声道系统中都有自己独特的发音方式, 因此不同音素产生爆破音的概率是不同的。为了提高爆破声的定位精度, 该防御方案仅在高概率音素存在的情况下将潜在的爆破音识别为真实的爆破音。尽管扬声器和麦克风通道的噪声引入导致了某些语音攻击样本也表现出爆破音, 但通过进一步地分析流行噪声的特征, 可以将真实人类(合法用户)发出的语音和电子设备重播的语音准确地区分开。

综上, 使用爆破音分析的方法可以有效地检测出大多数语音对抗样本。同时, Zhou 等人[34]还评估了不同类型的语音攻击(例如 DolphinAttack[35] 和 CommanderSong[26]), 使用不同长度的语音样本和通过不同扬声器播放语音的条件下该防御方案的有效性和鲁棒性。

4.1.2 样本分类器

除了通过语音活性检测来区分正常样本和对抗样本之外, 还有一种常用的方法是使用样本分类器来对不同的样本进行区分, 达到检测语音对抗样本的目标。基于机器学习技术的分类器可以通过分类结果检测出语音对抗样本并将它们隔绝在语音识别系统之前。因此, Carlini 等人[36]在语音识别系统中使用逻辑回归构造了一个隐藏的语音样本分类器。该分类器从语音的短期特征中提取的中期特征(例如均值和标准差)作为分类依据。实验表明, 通过将语音对抗样本标记为恶意样本, 该分类器最高可以实现 99.8%的恶意样本检测率, 而仅产生 0.2%的误报率, 这意味着该分类器只会错误地丢弃千分之二的正常语音样本, 对语音识别系统的性能影响几乎可忽略。对于一些基于白盒攻击精心构造的语音对抗样本, 该分类器的性能也能达到 70%以上的语音对抗样本检测率, 同时保证误报率小于 1%。因此, 使用分类器进行对抗样本检测具有一定的可行性。

由于现有的语音识别系统具有多样性, 它们使用不同的模型结构, 参数和训练数据集, 因此针对语音识别系统设计的语音对抗样本存在明显差异, 因此攻击的可迁移性较差。受到这一现象和多版本程序设计的启发, Zeng 等人[37]提出了一种新颖的语音对抗样本检测方法 MVP-EARS, 该方法通过分析各种针对语音识别系统生成的语音对抗样本来确定被检测语音是否为对抗样本。他们基于多版本程序设计类似的思路, 通过多种目标语音识别系统生成语音对抗样本, 建立了当时最大的语音对抗样本数据集。通过进一步调整上述思路, 该方法对语音对抗样本检测模型进行了主动训练, 对检测模型性能的评估表明 MVP-EARS 针对语音对抗样本的检测精度最高可达到 99.88%。MVP-EARS 可以大大降低了攻击者生成语音对抗样本的灵活性, 能够对语音对抗验本提供有效的防御。

与上述两个研究类似, Kwak 等人[38]也提出了一种新的基于文本转语音指令分析的用户合法指令识别和可疑语音指令检测的方法, 并使用一个大型的真实世界的语音系统数据集(大约 3460 万个语音, 包含 460 万个异常)评估其可行性。该方法创新地使用文本转语音指令语句和匹配的应用程序作为主要分类特征, 基于轻量级分类算法实现了对可疑语音指令的检测。为了评估检测准确性, 他们使用真实世界的语音系统数据集, 并测量了平均等错误率、检测准确率等指标。准确率验证的结果表明, 平均等错误率约为 3.4%, 检测准确率为 95.7%。

4.1.3 数据特征分析

Yang 等人[39]提出了一种根据时间依赖(temporal dependency)这一特定数据属性来对语音对抗样本进行有效检测的方法。该方法通过对三种先进的语音对抗样本生成技术所产生的语音对抗样本数据进行时间依赖性分析, 通过该特征可以在自适应和非自适应的攻击中有效的检测出语音对抗样本。与此同时, Yang 等人[39]还讨论了模型输入变换作为图像领域常用的降低对抗样本功能的常用技术在语音识别系统中的防御能力存在局限性。为了评估输入变换方法对语音对抗样本的防御能力, 他们在语音识别系统上实现四种输入转换方法, 分别是波形量化、时间平滑、下采样和自动编码器改造。对于最新的语音对抗样本攻击, 输入变换方法能够提供一定的的防御能力, 但防御效果有限。同时, 通过样本的特定数据属性来进行异常样本检测的思路不仅适用于语音领域, 也可迁移到其他研究领域用以提升模型的鲁棒性。

4.2 对抗训练

对抗训练的核心思想是将对抗样本添加到训练数据中来增强模型的鲁棒性, 进而提升模型对对抗样本的防御能力。不同于样本分类器的防御策略, 对抗训练策略将对抗样本作为训练过程的一部分, 以此来提升模型本身对该类攻击的防御能力。具体来说, 对抗训练过程先通过已知的多种对抗样本生成方法来获取大量的对抗样本, 然后将这些对抗样本添加原样本的标签并混入训练集, 这可以保证对抗扰动特征分布已经被包含在样本特征空间之中。然后, 随着在模型训练过程中不断加入拥有正确标签的对抗样本, 模型能够学习更泛化的样本特征分布。最终, 模型能够更准确地获取样本特征空间的分类边界, 从而使得对原样本进行微小扰动而产生的对抗样本失去对该模型的攻击能力。与此同时, 对抗训练可以通过增量学习的形式对现有的模型进行防御能力的提升, 这降低了对抗训练的使用成本, 提升了该防御方法的实用价值。

现有研究表明, 对语音识别系统进行对抗训练是一种防御已知语音对抗样本攻击的有效方法。Szegedy 等人[40]的研究工作证明将对抗样本和普通样本一起加入训练集来训练模型能够使该模型稍微正则化。增加对抗样本与增加训练数据并不一样, 其思路并不是单纯地扩充训练数据。增加普通样本会提升模型本身的性能, 而增加对抗样本并不能显著提高模型的性能, 但是可以揭露出模型的缺陷, 从而提升模型的鲁棒性。

由于 Szegedy 等人[40]在图像领域验证了对抗训练作为一种防御方法的可行性和有效性, Yakura 等人[27]讨论了对抗训练方法在语音领域的可能性。他们通过语音对抗样本生成技术获取的大量对抗样本, 然后将这些样本添加到训练数据中并赋予原语音的标签。此时数据集中同时包含正常语音样本和语音对抗样本, 在训练过程中, 语音识别模型被引导去学习对抗扰动的分布。最终, 模型通过提升自身鲁棒性获得了对语音对抗样本的防御能力。

不只是语音对抗样本可以进行对抗训练, 由于语音通过空气信道传播的特殊性, 语音重播样本也可以被用来提升模型的防御能力。从这个角度出发, Gong 等人[41]提出了语音数据集 ReMASC, 该语音数据集是为研究语音控制系统的漏洞和防御方案而提出的。ReMASC 数据集既包含真实的语音样本, 又包含每个语音通过空气信道重播后录制的重播样本, 这些样本可以用来训练数据集来提升模型对语音对抗样本的敏感性, 一定程度上可以辅助语音识别系统提升针对语音对抗样本攻击的防御能力。但是, 对抗训练只能对模型的安全性提供有限的提升。

4.3 数据压缩防御

4.3.1 音频压缩

数据压缩是一种应用起来比较容易的一种对抗样本防御方式。在图像领域, 已经有工作证明压缩技术能够减轻对抗样本带来的干扰。Das 等人[42]设计并实现了 ADAGIO 工具, 该工具可以进行对抗性音频攻击和防御的实验。在实验中 AMR 压缩和 MP3 压缩将针对性攻击的成功率从 92.5%降低到 0%, 并显著地提高了模型对于普通样本的正确率, 而且由于 AMR 基于心理声学原理设计的, 所以 AMR 压缩技术可以有效地从音频中去除人类无法感知的对抗成分。

4.3.2 降低采样率

降低音频采样率也是数据压缩防御方法中的一种, 它们都是通过修改输入语音以衰减添加到原始音频中精心制作的对抗扰动来破坏语音对抗样本的攻击能力。在 Yuan 等人[26]的研究工作中, 如果将音频采样率设定为 8000Hz, 下行采样率设定为 5600Hz, 可以将语音对抗验本的攻击成功率降低到 8%, 并且原有的语音仍然保持 91%的识别成功率。实验表明降低音频压缩率可以有效防御对抗样本攻击。但是数据压缩以及降低音频采样率不可避免地带来对原语音特征和信息的损坏, 因此是一把双刃剑, 需要在保持原语音和破坏对抗扰动间进行权衡。

4.4 模型优化

4.4.1 模型脆弱性分析

通过模型脆弱性分析来找出模型的缺陷是进行模型优化的一个重要步骤。Du 等人[43]提出了一种针对递归神经网络的模糊测试框架。递归神经网络(recurrent neural network, RNN)实现了时间行为和带有循环的“内存” 和内部状态, RNN 的这种有状态性质有助于其成功处理诸如音频, 自然语言和视频处理之类的顺序输入。作为 RNN 的典型应用, 语音识别系统面临测试不足的严峻问题。他们针对基于 RNN 系统提出了一种覆盖率指导的自动测试框架。考虑到RNN的独特网络结构, 他们首先将RNN模型形式化建模为马尔可夫决策过程, 并基于马尔可夫决策过程模型, 设计了一套专门针对基于 RNN 系统的测试标准, 以捕获其深度学习内部的动态状态转换行为。他们进一步提出了一个基于 RNN 系统的自动化测试框架, 该框架以指定的覆盖范围为指导。实验结合了 8 个音频变换方式以生成新的音频测试输入。基于巨大的测试生成空间, 实验利用覆盖率反馈来指导测试方向, 从而系统地覆盖 RNN 的主要功能行为和极端情况。实验从训练集中随机选择样本, 并大致遵循训练数据的分布以确保测试的有效性。

Zhang 等人[44]的工作也使用了模糊测试模型对自然语言处理中的意图分类器进行脆弱性分析。他们首先分析了语音助手行为中语音解释的语义不一致的问题, 并发现此问题是由意图分类器产生的。他们通过实验证明使用一些常见的口头错误时, 由意图分类器的不正确的语义解释导致的语义不一致会破坏语音助手处理事务完整性。为了对这种问题进行更进一步的分析, 他们设计了一个语言模型指导的模糊工具 LipFuzzer 来评估意图分类器的安全性, LipFuzzer 可以发现潜在的易于误解的口语错误。为了指导模糊测试, 借助统计关系学习和自然语言处理技术来构建对抗性语言模型。这种建模过程能够将语言知识转换为计算统计关系模型, 并最终通过该知识来分析模型的脆弱性问题。

4.4.2 防御性蒸馏

最初, Hintion 等人[45]提出了模型蒸馏方法, 该方法旨在将训练好的复杂模型具备的“知识”迁移到一个结构更为简单的网络中, 或者通过简单的网络去学习复杂模型中“知识”, 进而提升模型学习效率。通过对该方法稍加改动, 就形成了防御性蒸馏。一般来说, 复杂模型称为“教师模型”而相对简单网络称为“蒸馏模型”。

防御性蒸馏基于模型蒸馏方法来提高深度学习模型的稳健性, 但是与原蒸馏方法相比有两个显著的改变。第一点是教师模型和蒸馏模型在大小上是相同的, 换句话说, 防御性蒸馏不会对模型进行精简。第二点是防御性蒸馏使用一个更大的蒸馏温度常数来迫使蒸馏模型对其预测概率有更高的置信度。蒸馏过程可以降低计算对抗扰动时的梯度值并提高了生成对抗样本所需最小对抗扰动的平均值, 因此具有提高对抗样本生成难度的作用。以此为理论依据, Papernot 等人[46]提出使用防御性蒸馏来防御对抗样本攻击, 并通过实验验证了其作为防御方案的可行性。尽管防御性蒸馏能够提升语音对抗样本攻击的执行成本, 但只要稍加改进对抗样本生成技术就可以轻松绕过该防御, 同时对于黑盒攻击来说并不依赖梯度值来计算对抗扰动。

5 未来研究方向与挑战

综上所述, 如今语音对抗样本的攻防研究已经受到了全世界研究人员的广泛关注。针对语音对抗扰动的生成原理及其攻击方式的研究是一个重点领域。通过分析语音对抗样本攻击的特征, 提出系统而有效的语音对抗样本防御方案将是该领域的主要目标。作为一个年轻而又迅速发展的领域, 语音对抗样本未来可能会在以下几个方面获得进一步的研究和发展:

(1) 目前先进的语音对抗样本生成方法大多还是针对白盒模型开发的, 尽管攻击成功率高, 但是在真实场景下难以实现。如何实现更好的黑盒攻击, 如何提升黑盒攻击的效率, 以便生成更具有实用价值的语音对抗样本。

(2) 语音不同于图像, 它在通过空气信道时会不可控的引入大量的环境噪声, 严重破坏对抗扰动的攻击能力。尽管图像对抗样本攻击路径上也会引入由光线变化产生的噪声, 但这一问题在语音上要严峻的多。因此如何提升语音对抗样本对环境噪声的鲁棒性还需要进一步研究, 目前主流的方法是对环境噪声建模并在计算对抗扰动时考虑该因素的影响, 但是目前对环境噪声的建模能力还远不足, 无法应对复杂多变的真实物理世界。

(3) 由于人耳对语音的高敏感性, 使得针对人耳专门设计更具有隐蔽性的对抗扰动很有必要。目前主要的方法是利用心理声学原理来选择性的在原语音中添加对抗扰动, 尽管该方法被证明有效, 但是它约束了扰动的选择性, 使得语音对抗样本缺乏多样性, 更容易被检测。因此完善语音对抗样本攻击的隐蔽性具有很高的研究价值。

(4) 对抗扰动的生成方法不断革新, 相应的防御策略研究却进展缓慢。目前还缺乏针对语音对抗样本攻击全面且有效的防御方案。通过分析最先进的语音对抗样本攻击方法, 设计一个更完善更通用的防御方案来保障语音识别系统的安全性和可靠性具有重要的研究潜力和应用价值, 因此这一研究方向也是该研究领域的最终目标。

6 结语

最近的研究表明, 深度学习模型很容易被对抗样本误导而丧失识别能力, 这极大地威胁了依赖于深度学习模型的先进语音识别系统。在本文中, 我们介绍了语音对抗样本生成及其攻击方式的相关概念, 较为系统的总结了目前的语音对抗样本攻击方法, 讨论了相应的各类防御策略。

语音对抗样本是一种针对语音系统可用性的攻击, 该攻击可以干扰语音识别系统的正常功能或误导目标系统执行攻击者预期的恶意指令。基于攻击类型的差异, 我们将现有的语音对抗攻击分为两个主要类别; 根据防御角度的差异, 我们将现有语音对抗样本的防御策略分为四个类别。我们介绍了语音对抗样本攻击的基本原理并详细地回顾了最新的相关研究进展, 并且进行了全面的比较。最后, 本文从语音对抗样本攻击和防御两个方面讨论了未来的研究方向和挑战。

综上所述, 研究针对语音识别系统的对抗样本攻击和防御技术, 给出具有实际意义的抽象表达, 开发具有普适能力的对抗扰动计算方法和设计一个全面而有效的语音对抗样本防御方案, 实现安全高效的语音识别系统是该领域研究的最终目标。但是, 还需要前赴后继的研究工作来实现这一目标。

参考文献

[1] Wrench A A, Hardcastle W J. A Multichannel Articulatory Database and Its Application for Automatic Speech Recognition[C]. *th Seminar of Speech Production* , 2000:305-308.

[2] Povey D, Ghoshal A, Boulianne G, et al. The Kaldi speech recognition toolkit[C]. *IEEE 2011 workshop on automatic speech recognition and understanding (No. CONF)*, 2011.

[3] Huang X D, Alleva F, Hon H W, et al. The SPHINX-II Speech Recognition System: An Overview[J]. *Computer Speech & Language*, 1993, 7(2): 137-148.

[4] Hannun A, Case C, Casper J, et al. Deep Speech: Scaling up End-to-End Speech Recognition[EB/OL]. 2014: arXiv: 1412.5567. https://arxiv.org/abs/1412.5567

[5] Muda L, Begam M, Elamvazuthi I. Voice Recognition Algorithms Using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques[EB/OL]. 2010: arXiv: 1003.4083. https://arxiv.org/abs/1003.4083

[6] Itakura F. Line Spectrum Representation of Linear Predictor Coefficients of Speech Signals[J]. *The Journal of the Acoustical Society of America*, 1975, 57(S1): S35.

[7] LeCun Y, Bengio Y, Hinton G. Deep Learning[J]. *Nature*, 2015, 521(7553): 436-444.

[8] Kurakin A, Goodfellow I, Bengio S. Adversarial Examples in the Physical World[EB/OL]. 2016: arXiv: 1607.02533. https://arxiv.org/abs/1607.02533

[9] Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: A Simple Way to Prevent Neural Networks from Overfitting[J]. *Journal of Machine Learning Research*, 2014, 15(1): 1929-1958.

[10] Goodfellow I J, Shlens J, Szegedy C. Explaining and Harnessing Adversarial Examples[EB/OL]. 2014: arXiv: 1412.6572. https://arxiv.org/abs/1412.6572

[11] Hu S S, Shang X C, Qin Z, et al. Adversarial Examples for Automatic Speech Recognition: Attacks and Countermeasures[J]. *IEEE Communications Magazine*, 2019, 57(10): 120-126.

[12] Biggio B, Corona I, Maiorca D, et al. Evasion Attacks Against Machine Learning at Test Time[C]. *The 2013th European Conference on Machine Learning and Knowledge Discovery in Databases - Volume Part III*, 2013: 387-402.

[13] Amodei D, Anubhai R, Battenberg E, et al. Deep Speech 2: End-to-End Speech Recognition in English and Mandarin[EB/OL]. 2015: arXiv: 1512.02595. https://arxiv.org/abs/1512.02595

[14] Bispham, Mary K., Ioannis Agrafiotis, and Michael Goldsmith, A taxonomy of attacks via the speech interface[C], *Third International Conference on Cyber-Technologies and Cyber- Systems*, 2018:1-8.

[15] Wei X X, Liang S Y, Chen N, et al. Transferable Adversarial Attacks for Image and Video Object Detection[EB/OL]. 2018: arXiv: 1811.12641. https://arxiv.org/abs/1811.12641

[16] Gong Y, Poellabauer C. Crafting Adversarial Examples for Speech Paralinguistics Applications[EB/OL]. 2017: arXiv: 1711.03280.

[17] Kreuk F, Adi Y, Cisse M, et al. Fooling end-to-end speaker verification with adversarial examples[C]. *2018 IEEE International Conference on Acoustics*, *Speech and Signal Processing*, 2018: 1962-1966.

[18] Cisse M, Adi Y, Neverova N, et al. Houdini: Fooling Deep Structured Prediction Models[EB/OL]. 2017: arXiv: 1707.05373. https://arxiv.org/abs/1707.05373

[19] 2008 IEEE international conference on acoustics, speech, and signal processing (ICASSP)[C]. *IEEE Transactions on Audio*, *Speech*, *and Language Processing*, 2007: 1737.

[20] Carlini N, Wagner D, Communication N A B T, et al. Audio adversarial examples: Targeted attacks on speech-to-text[C]. *2018 IEEE Security and Privacy Workshops*, 2018: 1-7.

[21] Harik G R, Lobo F G, Goldberg D E. The Compact Genetic Algorithm[J]. *IEEE Transactions on Evolutionary Computation*, 1999, 3(4): 287-297.

[22] Alzantot M, Balaji B, Srivastava M. Did You Hear That? Adversarial Examples Against Automatic Speech Recognition[EB/OL]. 2018: arXiv: 1801.00554. https://arxiv. org/abs/1801.00554

[23] Taori R, Kamsetty A, Chu B, et al. Targeted adversarial examples for black box audio systems[C]. *2019 IEEE Security and Privacy Workshops*, 2019: 15-20.

[24] Neekhara P, Hussain S, Pandey P, et al. Universal adversarial perturbations for speech recognition systems[C]. *Interspeech 2019*, 2019: 481-485.

[25] Abdoli S, Hafemann L G, Rony J, et al. Universal Adversarial Audio Perturbations[EB/OL]. 2019: arXiv: 1908.03173. https://arxiv.org/abs/1908.03173

[26] Yuan X J, Chen Y X, Zhao Y, et al. Commandersong: A Systematic Approach for Practical Adversarial Voice Recognition[C]. *The 27th USENIX Conference on Security Symposium*, 2018: 49-64.

[27] Yakura H, Sakuma J. Robust Audio Adversarial Example for a Physical Attack[EB/OL]. 2018: arXiv: 1810.11793. https://arxiv.org/abs/1810.11793

[28] Qin Y, Carlini N, Goodfellow I, et al. Imperceptible, Robust, and Targeted Adversarial Examples for Automatic Speech Recognition[EB/OL]. 2019: arXiv: 1903.10346. https://arxiv.org/abs/1903.10346

[29] Shen J, Nguyen P, Wu Y H, et al. Lingvo: A Modular and Scalable Framework for Sequence-to-Sequence Modeling[EB/OL]. 2019: arXiv: 1902.08295. https://arxiv.org/abs/1902.08295

[30] Szurley J, Kolter J Z. Perceptual Based Adversarial Audio Attacks[EB/OL]. 2019: arXiv: 1906.06355. https://arxiv.org/abs/1906.06355

[31] Schönherr L, Eisenhofer T, Zeiler S, et al. Imperio: Robust Over-the-Air Adversarial Examples for Automatic Speech Recognition Systems[EB/OL]. 2019: arXiv: 1908.01551. https://arxiv.org/abs/1908.01551

[32] Chen Y X, Yuan X J, Zhang J S, et al. Devil's Whisper: A General Approach for Physical Adversarial Attacks Against Commercial Black-Box Speech Recognition Devices[C]. *The 29th USENIX Conference on Security Symposium*, 2020: 2667-2684.

[33] Abdullah H, Garcia W, Peeters C, et al. Practical Hidden Voice Attacks Against Speech and Speaker Recognition Systems[EB/OL]. 2019: arXiv: 1904.05734. https://arxiv.org/abs/1904.05734

[34] Zhou M, Qin Z, Lin X, et al. Hidden Voice Commands: Attacks and Defenses on the VCS of Autonomous Driving Cars[J]. *IEEE Wireless Communications*, 2019, 26(5): 128-133.

[35] Zhang G M, Yan C, Ji X Y, et al. DolphinAttack: inaudible voice commands[C]. *The 2017 ACM SIGSAC Conference on Computer and Communications Security*, 2017: 103-117.

[36] Nicholas Carlini, Pratyush Mishra, Tavish Vaidya, et al. Hidden voice commands[C]. *25th {USENIX} Security Symposium*, 2016: 513-530.

[37] Zeng Q, Su J H, Fu C L, et al. A multiversion programming inspired approach to detecting audio adversarial examples[C]. *2019 49th Annual IEEE/IFIP International Conference on Dependable Systems and Networks*, 2019: 39-51.

[38] Kwak I Y, Huh J H, Han S T, et al. Voice Presentation Attack Detection through Text-Converted Voice Command Analysis[C]. *The 2019 CHI Conference on Human Factors in Computing Systems*, 2019: 1-12.

[39] Yang Z L, Li B, Chen P Y, et al. Characterizing Audio Adversarial Examples Using Temporal Dependency[EB/OL]. 2018: arXiv: 1809.10875. https://arxiv.org/abs/1809.10875

[40] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing Properties of Neural Networks[EB/OL]. 2013: arXiv: 1312.6199. https://arxiv.org/abs/1312.6199

[41] Gong Y, Yang J, Huber J, et al. ReMASC: realistic replay attack corpus for voice controlled systems[C]. *Interspeech 2019*, 2019: 2355-2359.

[42] Das N, Shanbhogue M, Chen S T, et al. ADAGIO: Interactive Experimentation with Adversarial Attack and Defense for Audio[M]. Machine Learning and Knowledge Discovery in Databases. Cham: Springer International Publishing, 2019: 677-681.

[43] Du X N, Xie X F, Li Y, et al. DeepCruiser: Automated Guided Testing for Stateful Deep Learning Systems[EB/OL]. 2018: arXiv: 1812.05339. https://arxiv.org/abs/1812.05339

[44] Zhang Y Y, Xu L, Mendoza A, et al. Life after speech recognition: Fuzzing semantic misinterpretation for voice assistant applications[C]. *Proceedings 2019 Network and Distributed System Security Symposium*, 2019: 1-15.

[45] Hinton G, Vinyals O, Dean J. Distilling the Knowledge in a Neural Network[EB/OL]. 2015: arXiv: 1503.02531. https://arxiv.org/abs/1503.02531

[46] Papernot N, McDaniel P, Wu X, et al. Distillation as a defense to adversarial perturbations against deep neural networks[C]. *2016 IEEE Symposium on Security and Privacy*, 2016: 582-597.

台建玮 于 2016 年在北京邮电大学物联网工程专业获得学士学位, 于 2021 年在中国科学院大学网络空间安全获得博士学位, 现工作于合肥工业大学。研究方向包括: 人工智能应用技术, 深度学习可靠性。Email: 839066191@qq.com

李亚凯 于 2019 年在河北大学计算机科学与技术专业获得学士学位。现在中国科学院大学网络空间安全专业攻读硕士学位。研究兴趣包括: 人工智能安全, 深度学习可解释性。Email: liyakai@iie.ac.cn

贾晓启 于 2010 年在中国科学院研究生院信息安全专业获得博士学位。现任中国科学院信息工程研究所研究员。研究领域为系统安全。Email: jiaxiaoqi@iie.ac.cn

黄庆佳 博士学位, 高级工程师。主要从事操作系统安全、云计算安全、高级威胁检测、恶意代码分析等相关研究工作。Email: huangqingjia@iie.ac.cn