广州大学第十四届“挑战杯”

创业计划立项申报书

作品名称：一种针对语言中音色转换脱敏模型的攻击方法

申报者姓名：张景致

所在单位：广州大学 计算机科学与网络工程学院

申请日期：2023.3.12

项目分组：

R 科技创新和未来产业

□乡村振兴和产业发展

□ 城市治理和社会服务

□ 生态环保和可持续发展

□ 文化创意和区域合作

共青团广州大学委员会制

二○二三年三月

**说 明**

1.申报者在申报书封面上的项目类别上注意用“√”标明。

2.表内项目填写一律用钢笔或打印，字迹要端正、清楚。

3.申报书需一式两份（外加电子文档），并上交校团委。

**申报表**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 负责人情况 | 姓名 | | 张景致 | | | | | 性别 | 男 | | 出生年月 | 2003年1月 |
| 现学历 | | 本科 | | | | | | | | 联系电话 | 13660605355 |
| 专业 | | 软件工程 | | | 年级 | | 2 | 学制 | 4年 | 入学时间 | 2021年9月 |
| 申报项目全称 | | | | | 针对语言中音色转换脱敏模型的攻击方法的研究 | | | | | | |
| 项目所属领域分组 | | | | | | □A. ✓科技创新和未来产业 □B.乡村振兴和产业发展 □C.城市治理和社会服务 □D.生态环保和可持续发展 □E.文化创意和区域合作 | | | | | | |
| 是否设立实体 | | | | | |  | | 设立实体的名称 | | |  | |
| 所有团队成员  （最多10人） | | 姓名 | | | 性别 | | 学院 | 年级、专业 | 项目分工 | | | 联系电话 |
| 黄海 | | | 男 | | 计算机科学与网络工程学院 | 人工智能专业本科2年级 | 研究工作、数据分析、协助论文写作 | | | 13229368118 |
| 张景致 | | | 男 | | 计算机科学与网络工程学院 | 软件工程  本科2年级 | 协调工作、研究工作、撰写论文 | | | 13660605355 |
| 李菲菲 | | | 女 | | 人文学院 | 汉语言文学（师范）  本科2年级 | 财务预估和分析，商业价值评估，后期的宣传推广统筹和策划，商业准备 | | | 15814455773 |
| 冯琼英 | | | 女 | | 人文学院 | 汉语言文学（师范）  本科2年级 | 项目筹备，资料查阅总结，攻击模型的社会价值分析，文书撰写与润色 | | | 14767655499 |
| 徐燕 | | | 女 | | 人文学院 | 汉语言文学（师范）  本科2年级 | 项目简介，美工设计与宣传，项目策划书修缮 | | | 18138341895 |
|  | | |  | |  |  |  | | |  |
|  | | |  | |  |  |  | | |  |
|  | | |  | |  |  |  | | |  |
|  | | |  | |  |  |  | | |  |
|  | | |  | |  |  |  | | |  |
| 指导老师信息  （最多3人） | | 姓名 | | | 性别 | | 学院 | 职称 | 职务 | | | 联系电话 |
| 闫红洋 | | | 女 | | 计算机科学与网络工程学院 | 副教授 |  | | | 15920341511 |
|  | | |  | |  |  |  | | |  |
|  | | |  | |  |  |  | | |  |
| 申请经费（元） | | | |  | | | | | | | | |
| 是否获得过其他渠道的资助及金额 | | | |  | | | | | | | | |
| 项目简介  （500字以内） | | | | 随着深度学习算法的不断发展，该算法在近年来逐步被应用至互联网隐私数据安全攻防领域。作为隐私数据安全的一个分支，语音隐私保护领域目前仍以传统算法为主，但也出现了一些基于深度学习的语音脱敏算法，然其实际安全系数，还需实验检验。  本项目以语音隐私保护领域基于深度学习的语音脱敏算法为检验对象，通过对其进行漏洞清查、自我攻击，总结提出系列攻击方案与对应防御对策，旨在揭示现行技术下基于深度学习算法训练所得的音频音色转换脱敏模型仍具较大不可靠性，同时为深度学习算法在该领域的进一步研究发展提供参考。 | | | | | | | | |
| 社会价值  （500字以内） | | | | 语音脱敏是指以隐藏说话人身份信息为目的，祛除或改变说话人语音的个性特征。近年来，随着智能语音交互技术和声纹认证产品的快速发展，语音脱敏被广泛应用至语音产品隐私保护中，同时也屡被不法分子用以实施违法犯罪行为，使得公民隐私保护受到严重威胁。高水平需求压力之下，深度学习算法凭借较传统算法更强的学习能力、更广阔的覆盖范围等优势进入到语音隐私保护领域，但问题仍在持续暴露，落地产品准确率存在较大提升空间。  本项目分析了目前基于深度学习算法的语音脱敏模型的问题所在和技术难点，为语音隐私保护的未来研究发展提供参考依据，进而为人工智能时代公民隐私保护研究助力。 | | | | | | | | |
| 计划实践过程  （500字以内） | | | | 1.方向构思:语音安全隐私的领域主要依托于音频转换脱敏模型，而脱敏模型的建设需要不断地经过检验才能证明自身的安全可靠。本项目计划以攻击音频转换敏模型为目标，为提供语音服务等软件开发商提供脱敏模型自测途径，以此提升自身防御机制。2.论文研读:收集与音频处理、深度学习、人工智能安全领域相关论文，挑选其中符合项目研究方向的论文进行精细化阅读，以此完成专业知识学习。3.分析原理:在精读论文的基础上，对相关论文实现代码复现，分析其中的原理和网络结构。  4.设计模型:选取其中一个音频脱敏技术，设计出具有针对性的攻击方法，并实现有效攻击(完成模型的训练)。  5.评估与调整:完成攻击模型的设计后，评估攻击模型的性能和效果，根据实践进行对其调整，以达到攻击性和隐蔽性的平衡。  6.论文撰写与发表: 撰写论文，并对论文进行润色。论文发表。6.商业计划: | | | | | | | | |
| 创新意义  （500字以内） | | | | 随着互联网的不断发展，语音隐私保护领域逐渐受到重视，目前学界在语音脱敏相关领域虽然获得巨大进展，但存在许多不足，无法成功解决隐私信息泄漏的风险。对于提供语音服务的软件开发商来说，提供安全性能高的音频转换脱敏模型对其自身隐私安全建设有着不可估量的作用。语音安全隐私的领域主要寄托于音频转换脱敏模型，而脱敏模型的建设需要不断地经过检验才能证明自身的安全可靠。本项目计划以攻击音频转换敏模型为目标，设计出与已有方案角度不同的攻击方法，有利于为提供语音服务等软件开发商提供脱敏模型自测途径，有利于提升自身防御机制。同时我们将会注意提升攻击模型的隐蔽性和数据恢复的完整性。 | | | | | | | | |
| 发展前景  （500字以内） | | | | 深度学习算法广泛应用至语音隐私保护是互联网数据安全攻防领域的发展趋向，未来将有大量基于深度学习算法的语音脱敏模型需要进行漏洞排查、自我攻击，完成网络安全预演。本项目通过持续观测深度学习算法在语音隐私保护领域的应用成果，后续提供系列攻击方案，可为语音隐私数据安全领域长期靶场的建立奠定雏形。 | | | | | | | | |
| 团队协作  （500字以内） | | | | 团队由五名学生组成，内部分工明确，各司其职，配合良好。  张景致作为项目负责人，统筹项目开展，确保项目顺利进行，参考深度学习中图像脱敏的攻击方式，将攻击部分设计为神经网络层嵌入原模型网扩框架，提出一种针对语音脱敏技术的攻击方案，并撰写项目论文；  黄海作为核心技术人员通过查阅多篇权威论文并结合时下研究热点，提出最为流行的深度学习研究框架，结合自动编码器VAE、对抗生成网络GAN等前沿技术进行研究，对脱敏模型进行日常训练并对脱敏后的音频进行内容提取和性能评估与改进，同时协助论文写作；  冯琼英主要负责项目前期筹备，资料查阅与总结，攻击模型的社会价值分析，文书撰写与润色等工作；  徐燕主要负责制作项目简介，并结合攻击模型的特点，进行美工设计与宣传，同时对项目策划书进行整体修缮；  李菲菲主要对项目进行过程中的财务输出情况进行预估和分析，并对攻击模型的商业价值进行评估和展望，同时对后期的宣传与推广工作进行统筹和策划，为攻击模型实际落地使用做好商业准备。 | | | | | | | | |
| 项目介绍材料  （本部分内容请作为附件附在申报书后） | | | | 见附件 1 | | | | | | | | |
| 其他相关证明材料 | | | |  | | | | | | | | |
| 指导老师  意见 | | | | 签章： | | | | | | | | |
| 学院团委  意见 | | | | 签章： | | | | | | | | |
| 校团委  意见 | | | | 签章： | | | | | | | | |

**经费预算**

|  |  |  |
| --- | --- | --- |
| 预算支出项目 | 经费（元） | 根据及理由 |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
| 合 计 |  | |

**附件1**

**内容摘要**

本项目语音隐私保护领域基于深度学习的语音脱敏算法为对象，从隐私安全性的角度出发，分析了基于深度学习语音脱敏模型的不可靠性，本项目设计出一种具有针对性常见的基于深度学习的音频脱敏处理方法的攻击方法，提升攻击方法的隐蔽性与恢复数据的完整性，为软件开发商的内部防御测试提供演练方法。

创业团队

**指导教师：**闫红洋副教授从事机器学习数据安全与隐私保护方向研究。目前为止，已发表高水平学术论文30余篇，包括CCF A类或SCI一区期刊，如IEEE TDSC、Science China Information、Sciences、Information Sciences等，其中两篇入选ESI高被引论文，两篇会议论文获得国际学术会议最佳论文奖如CSS2019、ProvSec 2021。

**学生团队：**本项目团队中两位成员为计算机学院创新班成员，其中一位是创新班科研小组组长，具备一定的科研能力，在机器学习、深度学习、人工智能安全领域有较好基础。另外三名人文学院成员都具有良好的文字工作能力，并且在商业规划、商务沟通、宣传策划、预估分析有着不俗的能力，为后续商业计划推动落地提供后发之力。

**创业项目工作基础条件**

**硬件条件:**

拥有70平方米左右的实验室作为实验场地，能够同时容纳小组同时参与实验研究。

拥有学院提供英伟达 V100 GPU，可以提供强大的算力支持。

**团队积累：**

本项目团队中两位成员为计算机学院创新班成员，其中一位是创新班科研小组组长，具备一定的科研能力，在机器学习、深度学习、人工智能安全领域有较好基础。

**师资条件：**

本项目指导老师闫红洋老师在人工智能安全、联邦学习领域有着丰富的经验，可以为项目的计划、开发和论文编写等方方面面提供支持。

**研究积累：**

本项目自2022年7月开始研究，在查阅多篇有关内容论文后，我们提出、修改和完善了多版攻击架构。目前，实验所需的多个深度学习模型已经训练完毕，已有一定的实验结果，同时阶段性成果论文初稿已撰写完成。

**项目市场、财务竞争**

近年来，随着智能语音交互技术和声纹认证产品的快速发展，语音脱敏被广泛应用于语音产品隐私保护中，同时也屡次被不法分子用以实施违法犯罪行为，使得公民隐私具有泄露的风险，而如今学界对语音隐私保护领域的研究成果较少，因此本项目针对目前常见的音频脱敏处理方法提出的攻击方式研究具有很强的时效性，可以为相关研究领域提供及时的帮助。

本项目基于深度学习学习研究框架，结合自动编码器VAE、对抗生成网络GAN等前沿技术进行研究，参考深度学习中图像脱敏的攻击方案，将攻击方案部分设计为神经网络层嵌入原模型网络框架，提出一种设计方案并进行模型训练，在语音脱敏等相关领域具有很强的前沿性，可以更新相关领域的研究方法。

本项目的研究成果通过插入具有隐蔽性、恢复效果良好的语音优化隐写模型，使用较为复杂的（类)自动编码器结构对频谱进行处理，缓解了时序数据出现的数据不吻合问题和Griffin Lim算法的缺点，提高攻击效果，取得更好的实验成果，更好地为语音脱敏技术的更新发展以及语音隐私保护的未来研究提供参考依据，进而为人工智能隐私保护研究助力，具有强大的科技性。

本项目的项目成果作为一种攻击模型，可以为基于深度神经网络的音频脱敏转换模型的软件开发商提供自证服务，同时可以为企业抵御外来恶意攻击提供较好的演练作用，从而促进音频脱敏技术的优化和更新，体现企业对于用户的隐私保护，提高企业口碑和荣誉，具有很强的商业潜力。

**实际操作可行性**

**技术可行性**

1、模型选用：为了导出可用可信的结果，我们选用基于StarGAN的高性能多对多语言角色转换模型StarGAN-VC作为被攻击的对象，并使用当下十分先进的基于卷积神经网络可微分STFT和反STFT层的“Hide and Speak”隐写模型。这些模型都使用Python语言编写而成，使用PyTorch深度学习框架描述自身、管理和运行。

2、数据准备：同样为了结果的质量，我们使用被很多其他研究者信赖并使用的VCC 2016语音转换比赛数据集。这个数据集带有10余种不同的说话者一一对应的一百余条语音样本，已经预先做好了裁剪、预处理、标签和格式转换，适用于我们的模型训练和分析。

3、 模型训练：利用librosa和NumPy预先整理、合并和剪裁VCC 2016数据集中被我们选定的样本，产生相应数据文件。使用在 Anaconda 包管理平台上的 PyTorch 框架，利用英伟达CUDA并行运算、异构计算技术提高训练和推理速度。

4、实验、数据分析：使用和训练时同样的配置，将预先整理的数据使用先前训练的模型进行处理，导出每一步骤的结果音频（脱敏的音频、隐写后的音频和恢复出来的音频），使用科大讯飞说话者声纹识别API量化转化效果，使用NumPy、pandas对结果进行整理。

**商业可行性**

1、随着智能语音交互技术和声纹认证产品的快速发展，语音脱敏被广泛应用至语音产品隐私保护中，同时也屡被不法份子用以实施违法犯罪行为，使得公民隐私保护受到严重威胁。在此情况下，相关企业对隐私保护技术水平的要求有所提高，内测需求缺口扩大。

2、接触使用基于深度神经网络的音频脱敏转换模型的软件开发商，介绍本项目的攻击方法，希冀能够与对方团队达成合作，为其模型提供自证服务。同时，本项目的攻击方法对抵御外来恶意攻击有着较好的演练作用。此举能够加强企业口碑与信誉，与本项目实现合作共赢。

**风险分析**

1、 研究技术外漏，造成更大的信息安全隐患。

2、 项目合法性需进一步申报审批。

3、 需求方不信任，商业推进受阻。