附件1：

广州大学大学生创新训练项目计划申报书

|  |  |  |  |
| --- | --- | --- | --- |
| 项目编号 |  | | |
| 项目名称 | 针对语言中音色转换脱敏模型的攻击方法的研究 | | |
| 项目负责人 | 张景致 | 联系电话 | 13660605355 |
| 所在学院 | 计算机科学与网络工程学院 | | |
| 学号 | 32106300004 | 专业班级 | 软件211 |
| 指导教师 | 闫红洋 | | |
| E-mail | 32106300004@e.gzhu.edu.cn | | |
| 申请日期 |  | | |
| 项目期限 | 一年 | | |

广州大学

#### 一、 基本情况

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 项目名称 | 针对语言中音色转换脱敏模型的攻击方法的研究 | | | | | | |
| 所属学科 | 学科一级门：工科   学科二级类：计算机类 | | | | | | |
| 项目来源 | ■ A、来源于科技前沿或产业一线、具有一定创新性的基础研究课题或有针对性的应用研究项目  □ B、具有创新性的设计或制作类项目  □ C、开放实验室、实习实训基地中探索的综合性、设计性、创新性实验与训练项目  注：鼓励面向粤港澳大湾区重大战略需求和学校发展需要的项目；鼓励面向新兴边缘学科研究和跨学科的交叉研究项目。 | | | | | | |
| 申请金额 | 2000元 | 项目期限 | 1年期 | 拟申报项目级别 | | 校级 | |
| 负责人 | 张景致 | 性别 | 男 | 民族 | 汉 | 出生年月 | 2003.01 |
| 学号 | 32106300004 | 联系电话 | 手机： 13660605355 | | | | |
| 指导教师 | 闫红洋 | 联系电话 | 手机： 15920341511 | | | | |
| 项目简介 | | 该项目是针对深度神经网络语音声纹脱敏模型的一种攻击手法框架，通过在受攻击的模型上附加音频隐写模型来达到在脱敏后的语音内藏匿原音频的信息，而后从其中恢复原音频的目的。 | | | | | |
| 负责人曾经参与科研的情况 | | 无 | | | | | |
| 指导教师承担科研课题情况 | | 国家自然科学基金，青年项目，62102107，分布式机器学习中的隐私计算关键技术研究，30万元，2022.01.01-2024.12.31，主持。 | | | | | |
| 指导教师对本项目的支持情况 | | 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CSS 2019、ProvSec 2021。  [1]Hongyang Yan(闫红洋), Anli Yan\*, Li Hu, Jiaming Liang, Haibo Hu, MTL-Leak: Privacy Risk Assessment in Multi-task Learning. IEEE Transactions on Dependable and Secure Computing, 2023, DOI: 10.1109/TDSC.2023.3247869.（CCF 推荐A类期刊）  [2]Hongyang Yan（闫红洋）, Shuhao Li, Yajie Wang, Yaoyuan Zhang, Kashif Sharif, Haibo Hu, Yuanzhang Li\*. “Membership Inference Attacks Against Deep Learning Models via Logits Distribution”, IEEE Transactions on Dependable and Secure Computing, 2022, DOI: 10.1109/TDSC.2022.3222880. （CCF 推荐A类期刊）  [3]Anli Yan, Hongyang Yan\*(闫红洋), Li Hu, Xiaozhang Liu\*, Teng Huang. “Holistic Implicit Factor Evaluation of Model Extraction Attacks”, IEEE Transactions on Dependable and Secure Computing, 2022, DOI: 10.1109/TDSC.2022.3231271 （CCF 推荐A类期刊） | | | | | |
| 项目组主要成员 | 姓名 | 学号 | 学院 | 专业班级 | 联系电话 | 项目分工 | |
| 张景致 | 32106300004 | 计算机科学与网络工程学院 | 软件211 | 13660605355 | 技术研发、工程开发、论文编写 | |
| 李菲菲 | 32112170100 | 人文学院 | 汉语211 | 15814455773 | 项目筹备、资料调查、文书工作 | |
| 冯琼英 | 32112170055 | 人文学院 | 汉语213 | 14767655499 | 项目筹备、资料调查、文书工作 | |
|  |  |  |  |  |  | |
|  |  |  |  |  |  | |
| 指导教师 | 姓名 | 工号 | 学院/单位 | 职称 | 联系电话 | 电子邮件 | |
| 闫红洋 | 205334 | 人工智能与区块链研究院 | 副高 | 15920341511 | hyang\_yan@gzhu.edu.cn | |

#### 二、 立项依据与研究内容（可加页）

|  |
| --- |
| （1）研究目的 随着深度学习技术的不断发展，该技术在隐私保护的应用也逐渐被发掘。目前被各种服务采用的语音隐私保护技术仍然以传统算法为主，基于深度学习的技术未被广泛应用。在互联网隐私数据安全攻防的领域中，存在一些基于深度学习的较为先进的语音脱敏算法，然而通过我们的实验发现，受到攻击的音频脱敏模型可以输出仍然能够恢复原有的隐私信息的语音，这意味着基于深度学习的语音脱敏算法中存在着严重的安全隐患。本项目提出一种针对目前常见的基于深度学习的音频脱敏处理方法的攻击方案，旨在揭示基于GAN的音频音色转换脱敏模型的潜在的不可靠性，并为改进这样的音频音色转换脱敏模型乃至语音脱敏、隐私保护技术的其他研究提供参考。 （2）研究内容 （此部分为重点阐述内容） 在互联网高速发展的时代，语音技术的频繁使用，承载着丰富隐私信息的音频面临着泄露的风险。通过声音的音色、频率等相关指数可以分析出语音发出者的身份、性别、年龄甚至是性格等隐私信息。出于对音频隐私信息的保护，基于深度学习的音频脱敏工具被研制出来，它们能实现保留传达内容的同时，过滤隐私信息。然而，对于其脱敏的完全性、可靠性我们保持怀疑态度。  本项目以基于深度神经网络的音频脱敏转换模型为攻击对象，研究出具有针对性的攻击方法，最终能达到恢复脱敏音频中隐私信息的效果。同时，我们也注意提高攻击模型的隐蔽性和数据恢复的完整性。最终希望通过攻击方法的设计，在预防攻击领域形成一定的认识，为将来此类防御工作的建构奠定基础。 （3）国、内外研究现状和发展动态 随着信息技术、网络通信等科学技术的发展，智能语音识别技术、声纹识别技术已经普及到我们生活的各个角落，更加先进的技术也不断涌现[7][8]，新兴技术成果为人们的生活带来便利的同时也引发了一系列的隐私泄露问题，从语音信息中泄露的隐私信息对用户的财产安全、人身安全造成威胁，针对以上问题，同时为了避免语音终端服务商以及其他潜在攻击者对用户语音的发掘与利用，相关领域的研究者试图打造一种保留音频内容而隐藏用户声音所携带的个人信息的服务工具，即对音频进行“脱敏”处理，为网络空间安全防范与建设助力。  通过查阅相关研究资料发现，许多研究学者从语音识别本质特性等方面入手，结合多种研究方式和识别算法提出多种脱敏技术方案，如张宁池等人提出一种通过构建包含资源层、数据层、服务层以及应用层的 B/S 的管理结构，在用户进行信息交互时，保护用户的敏感隐私信息的大数据脱敏技术方案 [1] ；王勇通过对生物特征模板、加密算法的研究提出三种基于感知哈希的多格式语音认证算法及两种基于生物哈希的多格式语音认证算法，并针对现存语音生物哈希检索系统中明文数据的泄漏问题，提高语音检索精度、语音特征多样性、可撤销性，提出一种基于特征融合的长序列生物哈希密文语音检索[3]；王雪瑞聚焦实现对声纹认证系统的活体检测研究并对目前已有的声纹活体检测方案存在问题进行总结，提出了一种安全高效、通用灵活的基于声纹认证系统的活体检测方案，使用随机语音挑战的形式，能够抵抗欺骗（重访和模拟）攻击，增强声纹认证的安全性，并开发了对应的手机软件。[2]语音脱敏技术主要针对于内容和声纹两方面。在声纹方面，基于风格转换模型StarGAN的StarGAN-VC模型在多对多说话者角色转换方面成绩较为突出[4]。  目前学界虽然在语音以及大数据信息脱敏相关的研究领域已经获得巨大进展，但依然存在许多不足，无法成功解决隐私泄漏的的问题，从整体而言，针对语音脱敏的全流程开展的研究较少，虽然仍有个别较为完善的方案[6]，但其余的基本针对音频内容或声纹的特点从局部展开研究，因此语音脱敏范围仍然需要进一步开拓；从脱敏的结果分析上看，学界目前对脱敏之后的结果分析方法比较匮乏，多数采用经验主义的分析方法，除此之外，学界对结果分析的方法尚没有广为认同的分析方法，从而导致没有很好的办法对脱敏模型的实际效果进行评估和改进，使得脱敏技术的整体进步受到限制。  源于上文所言的研究现状，我们的项目采用时下最流行的深度学习框架PyTorch，结合VAE、GAN等前沿技术进行研究，采用学术界前沿的隐写模型[5]，提出一种针对基于深度学习的语音脱敏技术的，与一些已有方案[9]角度不同的攻击方案，试图指出目前脱敏技术的不足，并为语音脱敏技术的进一步更新发展和隐私保护助力。    参考文献：  [1]张宁池,朱小娟,张宇等.互联网商业模式下大数据脱敏方法的探讨与研究[J].自动化技术与应用,2021,40(01):150-154.  [2]王雪瑞.一种具有活体检测功能的声纹认证系统[D].西安电子科技大学,  ,2021.DOI:10.27389/d.cnki.gxadu.2021.003585.  [3]王勇. 生物特征模板在语音认证与检索算法中的研究[D].西北师范大学,2021.DOI:10.27410/d.cnki.gxbfu.2021.001574.  [4] StarGAN-VC: Non-parallel many-to-many voice conversion with star generative adversarial networks  [5] Hide and Speak: Deep Neural Networks for Speech Steganography  [6] Overo: Sharing Private Audio Recordings  [7] Robust Speaker Recognition from Distant Speech under Real Reverberant Environments Using Speaker Embeddings  [8] LiVoAuth: Liveness Detection in Voiceprint Authentication with Random Challenges and Detection Modes  [9] Backdoor Attacks against Deep Neural Networks by Personalized Audio Steganography （4）创新点与项目特色 （此部分为重点阐述内容） 1. 在研究主题上，本项目关注时下颇受关注的互联网信息传播与隐私安全问题；在研究框架搭建上，本项目采用时下最流行的深度学习框架，结合自动编码器VAE、对抗生成网络GAN等前沿技术进行研究；  2. 在研究前沿性上，学界目前对语音编解码、合成和隐私保护领域的研究成果较少，本项目参考近3年深度学习和音频处理的多篇论文，基于CNN卷积神经网络、音频隐写技术、GAN生成式对抗网络等深度学习方法，同时为了增强攻击的隐蔽性，降低攻击受察觉的几率，提高实验效果，本项目将攻击部分设计为神经网络层嵌入原模型网络框架，提出一种针对于基于深度学习的语音脱敏技术的攻击方案；  3. 在研究意义上，本项目研究针对目前常见的音频脱敏处理方法进行设计攻击方案，指出当下音频脱敏处理的不足，旨在为语音脱敏技术的更新发展和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。 （5）技术路线、拟解决的问题 （此部分为重点阐述内容） 技术路线：   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。   拟解决的问题：   1. 采用的模型是对应特定的计算机硬件、软件配置而设计的，在我们不同于作者的配置下复现相应的结果存在困难，我们将深入分析采用的模型，进行必要的改进以令其正常工作。 2. VCC 2016数据集虽然是预处理完毕的，但若是直接使用磁盘上的音频源文件会造成IO瓶颈，需要对数据进行更多的整理以提高效率。  （6）预期成果及应用前景 （指项目成果能否在相关行业中推广、使用或解决行业中存在问题） 预期成果：   1. 发表SCI论文一篇 2. 完成人工智能安全相关专利申请一项   应用前景：   1. 在安全审计领域可作为一种部署、使用均较为方便的攻击手法使用 2. 在针对这类攻击的检测模型中作为参考模型使用，   **（7）项目研究进度安排 （指周期内阶段工作安排）**  **2022.7-2022.8** 研读与音频、深度学习、人工智能安全领域相关的论文，完成相关知识的学习。  **2022.9-2023.1** 在广泛研读相关论文后，精选其中符合研究课题的论文，对相关论文实现代码复现，并分析其中的原理和网络结构。  **2023.2-2023.8** 选取其中一个音频脱敏技术，设计出针对性的攻击方法并实现有效的攻击（完成模型的训练）。  **2023.9-2023.12在**完成攻击模型的设计后，评估攻击模型的性能和效果，根据实践对其进行调整，以达到攻击性和隐蔽性的平衡。  **2024.1-2024.3** 论文撰写。  **2024.4** 论文发表。 （8）已有基础  * 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一块，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 |

#### 三、 经费预算

| **开支科目** | **预算经费（元）** | **主要用途** |
| --- | --- | --- |
| 材料费 | 500 | ■实验材料，□图书，□办公耗材 |
| 业务费 | 1000 | □检测测试，■知识产权服务， □加工 |
| 印刷服务费 | 500 | ■打印复印，■版面费 |
| 其他 | 0 | 开支内容：无 |

#### 四、 申请人项目原创性承诺

|  |
| --- |
| 申请人承诺本项目为原创项目  **申请人签字：**  **年        月        日** |

#### 五、 指导教师项目原创性承诺及推荐意见

|  |
| --- |
| 通过    **指导教师（签章）：**  **年        月        日** |

#### 六、 院系推荐意见

|  |
| --- |
| **盖 章：**  **年        月        日** |

#### 七、 学校推荐意见

|  |
| --- |
| **盖 章：**  **年        月        日** |

|  |  |
| --- | --- |
|  |  |