**大学生科技创新项目**

**申 报 书**

**项目名称：**

**项目申报人：**

**学校名称：**

**申报日期：**

**项目类别：个人项目□ 团队项目□**

**浙江省大学生科技创新活动计划（新苗人才计划）实施办公室 制**填写说明

一、填写申报书前，请先查阅《浙江省大学生科技创新活动计划(新苗人才计划)实施办法》及申报通知。

二、申报书要按照要求，逐项认真填写，填写内容必须实事求是，表达明确、严谨。

三、格式要求：申报书中各项内容以Word文档格式填写，表格中的字体为**小四号仿宋体，1.5倍行距**；表格空间不足的，可扩展。

四、申报书由所在学校审查、签署意见并加盖公章后,报送浙江省大学生科技创新活动计划（新苗人才计划）实施办公室。

**一、项目简介**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 项  目  概  况 | 项目名称 | | |  | | | | | | | |
| 项目性质 | | | （）基础研究 （）应用基础研究 | | | | | | | |
| 项目来源 | | | （）自主立题 （）教师指导选题 | | | | | | | |
| 起止时间 | | | 自 2021年 3 月 至 2020 年 12 月 | | | | | | | |
| 项目状况 | | | （）研发阶段 （）中试阶段 （）批量（规模）生产 （选项打√） | | | | | | | | |
| 项  目  申报  人 | | 姓名 |  | | 性别 |  | 出生  年月 |  | 入学年月 |  | |
| 院系  专业 |  | | | 联系  电话 |  | | 电子  信箱 |  | |
| 项 目 组  主要成员 | | | 姓名 | | 联系电话 | | 院系专业 | | 年级 | | 具体分工 |
|  | |  | |  | |  | |  |
|  | |  | |  | |  | |  |
|  | |  | |  | |  | |  |
|  | |  | |  | |  | |  |
| 项  目  指  导  老  师 | | | 姓名 | | 联系电话 | | 所在单位 | | 职称 | | 主要研究方向 |
|  | |  | |  | |  | |  |
|  | |  | |  | |  | |  |
|  | |  | |  | |  | |  |
| 近三年成果：国家级 等奖 项，省部级 等奖 项 | | | | | | | | |
| 近三年科研经费 万元，年均 万元 | | | | | | | | |
| 项  目  主  要  内  容  简  介 | | |  | | | | | | | | |

**二、项目背景、目的及意义**

|  |
| --- |
| （简要说明项目背景、研究现状、意义及实施必要性）  1、**项目背景**  2017年7月20日，国务院发布了《新一代人工智能发展规划》，中国已进入人工智能的蓬勃发展期，以深度神经网络为代表的人工智能技术飞速发展，大幅跨越了科学与应用之间的“技术鸿沟”，诸如图像分类、语音识别、知识问答、人机对弈、无人驾驶等人工智能技术实现了从“不能用、不好用”到“可以用”的技术突破，迎来爆发式增长的新高潮。人工智能从1956年Dartmouth会议上正式提出，到如今经过60多年发展已经得到了巨大的发展与突破。  目前人工智能的基础与核心问题是机器学习，深度学习又是机器学习的一个重要领域与分支，其构建了一个模拟人类脑神经元的深度神经网络。2006年Hinton等人提出了深度学习的概念，之后几年，深度学习研究成为了人工智能领域的重大热点，其成功的应用领域包括计算机视觉、语音识别、推荐系统等等，深度神经网络已经在不同的应用场景下实现了非凡的成功。  然而，研究工作已经表明深度神经网络对于对抗样本这种输入是十分脆弱的。一名攻击者可以通过添加少量人为设计的扰动到原始的图像输入中，以此来构成对抗样本去误导深度神经网络为基础的应用。这些恶意生成带有微小扰动的实例可以造成学习模型系统将它们错误分类到攻击者事先设计好的目标类别（这种攻击称为有目标攻击）或者错误分类到与原来输入实例的真实标签类别不同的类别中（这种攻击称为无目标攻击）。  目前很多算法都被提出来用于生成这种对抗样本，如基于优化的方法和快速梯度符号标记法等等。大多数目前的攻击方法是基于最优化方程和简单的像素空间的矩阵度量。2014年Goodfellow等人提出了生成对抗网络（GAN），这是一种无监督学习方法，通过训练可以得到以假乱真的数据。GAN能够提高图像的质量，减少训练时间从而提高效率，多用在数据增强、风格迁移等应用中。GAN给深度学习的研究带来了新方向，当然也被运用在了对抗样本的生成上。  验证码（CAPTCHA）是“Completely Automated Public Turing test to tell Computers and Humans Apart”（全自动区分计算机和人类的图灵测试）的缩写，是一种区分用户是计算机还是人的公共全自动程序。其可以防止恶意破解密码、刷票等暴力行为，有效防止黑客对某一个特定注册用户用特定程序暴力破解方式进行不断的登陆尝试。目前验证码的载体形式多样，最常见的如图片验证码、语言验证码等。  随着图像识别技术的提高，简单的图片验证码能够被机器精准识别，因而黑客可以利用自动化验证码识别工具使验证码失去防护作用。图像验证码相对来说是简单且成本较低的安全防护程序，因此提高图像验证码被机器识别的能力具有巨大的意义和研究价值。将基于GAN生成的对抗样本应用于图像验证码，通过对抗样本的细微扰动，在不影响人类识别图像验证码的情况下尽可能让机器识别错误将是一个极具研究意义的领域。  **2、研究现状**  在2013年，Szegedy首次提出了通过对数据加入轻微扰动得到新样本能使特定的机器学习模型产生错误的决策分类，这种加入轻微扰动之后的新样本就是对抗样本。后来，有学者研究证明机器学习模型在处理对抗样本时会有漏洞，即对抗样本会成为训练机器学习算法的盲点。在Szegedy的研究基础上，研究者围绕不同的深度学习应用展开了对抗样本的研究。2015年，Goodfellow等人在ICLR2015上提出了一种有效的方法计算对抗扰动，即快速梯度标记法。2016年，Kurakin等人将手机摄像头拍摄到的对抗样本输入Inception分类器中，证明对抗样本在物理世界中存在潜在威胁。同样在自然语言处理中，替换单个词汇往往可以极大地改变文档的语义，如在2018年，Alzantot等仅通过少量词汇的变动就以97%和70%的成功率攻击了情感分析和文本蕴含模型。再者，2019年，Qin等构造了人耳无法辨别的音频对抗样本，通过对真实环境的模拟，展示出对抗样本对无线音频技术同样存在潜在威胁。可见，目前深度学习应用的各个领域，如计算机视觉、自然语言处理、语音识别等都会受到对抗样本的影响。  攻击方法根据攻击者对攻击模型的了解程度大致可分为三类：黑盒攻击、灰盒攻击和白盒攻击。1）黑盒攻击：攻击者通过输入获得相应的输出，并根据输出产生对抗样本，这种方式因为获取的数据最少，所以生成对抗样本的难度最大。黑盒攻击的方法有如下：2017年Athalye等人提出的期望攻击算法以及Chen等人提出了基于置信度的零阶优化方法，还有同年Brendel等人提出的仅需输出类型的决策边界攻击；2019年Su等人提出了一种基于差分进化算法的极少像素攻击，Brunner则提出了有偏的边界攻击，极大的提高了攻击的效率。2）灰盒攻击：攻击者不但能够获得输入输出，还能获得模型内部除了防御机制以外的结构和参数，这种攻击的对抗样本生成方式相对黑盒攻击较简单。3）白盒攻击：白盒攻击能够获得模型的一切信息，是最简单的对抗样本生成方式。针对白盒攻击建立的方法有如下：2013年Szegedy等人在提出对抗样本概念的同时也利用L-BFGS建立了模型；2014年Goodfellow等人提出第一个利用梯度的快速梯度下降算法FGSM；2016年，Kurakin等人在FGMS的基础上提出了基础迭代法（BIM），对FGMS进行了扩展；2017年Carlini和Wagner针对防御蒸馏算法提出了C&W算法，能够对蒸馏或者未蒸馏的模型都达到非常完美的攻击效果。  攻击方法根据是否有针对性目标可以分为定向攻击和非定向攻击：1）定向攻击：让分类模型将对抗样本分类为指定的类别；2）非定向攻击：可以让分类模型将对抗样本分类为除正确类别以外的任意类别，这种攻击相对简单。  防御对抗样本的方法主要分成三个类别：输入预处理、提高模型的鲁棒性和恶意检测。具体的方法有很多种。1）预处理的方法有以下几种：2017年Wang等人通过在样本中随机消除特征来阻止攻击者构建有效的对抗样本，同年Das提出一种能够搭建在已训练好的网络模型上的JPEG压缩预处理模块。2）提升模型鲁棒性的方法有如下：2014年Goodfellow首次提出了利用在模型训练中加入对抗样本的方法来提升模型的鲁棒性；2016年Papernot等人提出了防御蒸馏的方法，但是此方法完全无法防御C&W攻击；2019年Chang提出了一种基于双抗样本的训练方式，既能抵抗单步的对抗样本，也可以防御迭代的对抗样本。3）恶意检测方面：2017年Metzen等人提出增加一个小的检测子网络来扩充深度神经网络，这个子网络用于检测对抗样本的二分类器；2018年Tian等人提出了一种基于图像变换的对抗样本检测方法；2019年Chen等人首次提出基于有状态检测的黑盒防御方法，相比于目前研究的无状态防御，该防御增强了防御方的能力。虽然目前有多种多样的防御方法，但是并没有一种可以完全防御所有的对抗攻击和对抗样本。  **3、项目意义**  验证码作为一种区分正常人和机器操作的识别手段，可以有效防止网站被暴力破解，提高破解密码的难度。网站用户的注册、登录、发帖、投票等等应用场景，都有被机器刷的可能，造成各类损失。如果不对这类行为进行防范，灌水内容、垃圾注册、恶意登录、刷票、撞库等行为将大概率发生，一旦发生，将对产品自身发展和用户体验造成严重的影响。  目前网站常用的验证码依然是图片验证码，如数字等传统字符验证，这类验证码操作简单，减少了开发者的劳动量和网站的运营成本。随着深度学习的发展与应用，机器识别图片验证码是一件容易的事，被识别破解后就形同虚设了。  通常机器识别验证码的背后是一个机器学习的识别模型，基于前期大量的训练学习，能够精准地对图片验证码进行识别。而基于GAN生成的对抗样本图片验证码可以降低机器识别的精确度，甚至让机器识别失效，有效防止用户利用机器识别验证码而对网站进行恶意攻击。  **4、项目实施必要性**  互联网的飞速发展使网站建设达到顶峰，无论是新闻、购物、娱乐还是学习都离不开网站和互联网，通过特殊构造的图片验证码从而使机器无法正确识别验证码可以极大提高网站和用户信息的安全性。 |

**三、项目研究方案**

|  |
| --- |
| （包括项目的主要内容、计划目标、思路方法、组织实施、进度安排等）  **1、项目主要内容**  该项目主要进行了对抗样本生成方法的研究，提出了基于生成对抗网络（GAN）生成对抗样本的方法，提高了对指定目标类型的对抗攻击成功率，同时也提高了生成的样本图片的质量。运用生成的图片验证码对抗样本，提高图片验证码抵抗被机器正确识别的能力。  本项目的内容主要分为以下两个部分：  （1）基于GAN的对抗样本生成方法  对于对抗样本生成算法的研究之前主要集中在快速梯度攻击和基于优化的方法，这两种方法拥有各自不同的优势，其中快速梯度攻击使用的是快速梯度下降思想，提高了生成的效率，也提高了生成样本使分类器错误分类的概率，但是由于算法受限，已无法进一步提升。基于优化的方法生成对抗样本的效率较低，尽管对抗效果较好。之后，随着深度学习的发展，卷积神经网络迎来了研究和应用热潮。有研究者通过卷积神经网络生成对抗样本，但是图片质量不佳。随着GAN的提出，GAN成为了深度学习领域的研究热点之一。GAN主要有两个模型：Generator（生成器）和Discriminator（判别器）。Generator能够通过给定的输入生成特定的输出，如图片；Discriminator可以用来判别输入的图片真的还是假的，即是真实样本还是Generator生成的样本，其作用是使Generator生成的样本无限接近真实样本，直至Discriminator无法正确辨别输入图片的真假。基于GAN生成的对抗样本，不仅对抗样本效率高，且生成的图像逼真质量高。  本项目提出了一种基于AdvGAN模型的对抗样本生成模型（Gen-AdvGAN），在AdvGAN模型基础上进行改进。其核心思想是将干净样本通过GAN的生成器映射成对抗扰动，然后将该对抗扰动与相对应的干净样本相加，生成对抗样本，利用该生成的对抗样本且该对抗样本与原始干净样本肉眼几乎看不到差别，但原本能够正确识别判断干净样本的模型却无法正确判断该对抗样本。  下图1为AdvGAN的总体架构图，总共由三部分组成：生成器***G***，判别器***D***和目标神经网络***C***。将干净样本***x***输入到***G***中生成对抗扰动***G(x)***，然后将***G(x)***与***x***相加得到对抗样本***x’***，分别将***x’***和***x***发送给判别器***D***，用于区分生成的样本和原始的干净样本，判别器***D***的主要目的是鼓励生成的实例与原始类中数据不可区分。为了使目标神经网络***C***判断错误，再将生成的对抗样本***x’***输入到目标神经网络***C***中，其损失函数为***LADV***，表示预测与目标类别***t***目标攻击之间的距离。优化目标损失函数，当模型达到最优时，***G(x)***即为对抗扰动。损失函数由三部分组成：  其中***LGAN***为对抗损失，优化该损失函数的目的是为了鼓励生成的数据分布接近原始分类的数据分布。***LADV***为误判损失，优化该损失函数的目的是为了使模型往对抗样本的方向去生成。当G(x)达到最优，目标神网络C就能在一定程度上分辨出对抗样本，以此来增加模型C的鲁棒性。***Lhinge***为hinge损失，hinge损失用于正则化，稳定GAN的训练。    图1.AdvGAN架构图  本项目提出的基于GAN的对抗样本模型架构图如下图2所示。该模型与AdvGAN模型最大的不同是尽可能让误判损失***LADV***变大，但是又要让对抗样本和原始干净样本尽可能相似。在该架构中，对于被攻击的神经网络***C***，其有极高的精确度识别原始干净样本，但是对于生成的对抗样本，其会出现很低的精确度。    图2.Gen-AdvGAN架构图  （2）利用对抗样本实现验证码系统  基于（1）中得到的高质量的对抗样本，将之运用到实际应用当中。将由Gen-AdvGAN生成的对抗样本运用在验证码系统中，用对抗样本验证码图片代替原来的验证码图片，降低验证码系统被机器自动识别的风险，进而提高网站的安全性。该系统主要由两部分流程组成：用户行为流程和图像刷新流程。  （1）用户行为流程    图3.用户行为流程  如图3所示，用户进入验证码系统后，将得到刷新的图片验证码，如数字验证码图片，之后用户需要识别验证码图片中的数字然后输入判断框，若识别错误，则验证码图片刷新用户将重新识别输入；若识别正确，则验证通过。  （2）图像刷新流程  当用户输入验证码错误，后台随机从验证码图片库中取一张图片，再将该图片输入Gen-AdvGAN模型中，通过生成器得到一张对抗样本图片，并将该图片展现在网页前端。判断用户的输入是否与图片标签一致，若一致，则验证通过，反之，刷新图片，重复以上步骤。验证码图像刷新流程如下：    图4.图像刷新流程  **2、计划目标：**  完成项目后应达到的目标如下：  （1）对于特定的验证码图片识别模型，对本项目提出的Gen-AdvGAN网络训练生成的对抗样本的识别准确精度不高于40%。  （2）生成的对抗样本图片与原始干净样本的相似度达到95%，存在细微差异。  （3）提出的验证码系统拥有良好的用户体验，可以提高主流的基于深度学习的验证码识别器的识别次数。  **3、思路方法：**  通过研读相关论文，复现模型，对不同已有方案和模型进行实验与比较，归纳总结各模型的优点与缺点，开发改进适用本项目场景的模型。对提出的新模型进行调整与测试，通过大量实验进行模型优化。  **4、组织实施：**  **第一阶段：实现Gen-AdvGAN网络（2021.03-2021.08）**  吴炳金、周楠负责Gen-AdvGAN网络的搭建与训练；常一可负责模型的测试。  第二阶段：制造自定义的干净的验证码图片库（2021.08-2021.10）  第三阶段：调整网络模型，使模型适用于验证码图片库（2021.11-2022.4）  第四阶段：开发部署验证码系统（2022.5-2022.11）  **5、进度安排：**  **第一阶段（2021.01-2021.03）**  搜集整理资料，展开调研。  **第二阶段（2021.04-2021.09）**  基于相关知识展开实验。  **第三阶段（2021.10-2022.02）**  完成模型的训练与测试。  **第四阶段（2022.03-2022.05）**  开发部署系统。  **第五阶段（2022.06-2022.08）**  测试改进系统。  **第六阶段（2022.09-2022.12）**  整理创新点，撰写论文专利等，撰写解题报告。 |

**四、项目研究条件及创新之处**

|  |
| --- |
| （包括实施该项目所具备的基础、优势和风险，以及项目创新点等）  **1、实施项目所具备的基础、优势**  1）本项目成员年级分布为本科二年级、三年级和研究生一年级不等，其中吴炳金、周楠等成员有扎实的深度学习基础，能够高质量完成深度学习模型的开发与训练。  2）本项目指导老师  3）  **2、实施项目的风险**  1）实现该项目需要较深厚的深度学习理论知识与出色的实践动手能力，需要项目成员不断学习前沿的深度学习知识和模型。  2）发表该项目的论文需要极好的英语论文撰写能力和学术素养，项目成员需要花费很大精力学习撰写论文并发表论文。  **3、项目创新点**  1）本项目通过应用当前人工智能领域火热的生成对抗网络来生成对抗样本，对其他深度学习模型进行干扰与攻击，研究深度学习模型的学习盲点，这在人工智能安全领域具有重大意义。  2）将所研究的理论知识应用于实践，开发验证码系统，提高网络安全，将人工智能安全与网络安全相结合，为之后对安全的进一步研究打下坚实基础。 |

**五、项目预期成果**

|  |
| --- |
| （包括知识产权成果和经济、社会效益等）  **1、知识产权成果**  （1）根据项目需求，开发符合要求的验证码系统；  （2）申请发明专利1项；  （3）申请软件著作权1项；  （4）投递核心及以上期刊论文或论文成果。  2、项目的经济效益  3、项目的社会效益 |

**六、项目财务预算**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| （包括经费预算及经费支出明细等）  总计预算10,000元。   |  |  |  | | --- | --- | --- | | **序号** | **经费开支科目** | **预算金额（元）** | | 1 | 论文版面及专利申请费用 | 6,000 | | 2 | 专业书籍费用 | 1,000 | | 3 | 培训差旅费用 | 2,000 | | 4 | 打印材料费用 | 1,000 | | 5 | 合计 | 10,000 | |

**七、审核流程**

|  |  |
| --- | --- |
| 承诺书 | 1．本报告中所填写的各栏目内容真实，准确。  2．提供验收的技术文件和资料真实、可靠，技术（或理论）成果事实存在。  3．提供验收的实物（样品）与所提供鉴定的技术文件和资料一致，并事实存在。  4．本项目的知识产权或商业秘密明晰完整，未剽窃他人成果．未侵犯他人的知识产权或商业秘密。  5．项目实施经费合理有效，由承担项目的学生使用，无弄虚作假行为。  若发生与上述承诺相违背的事实，由项目组承担全部法律责任。  签名（全体成员）：  年 月 日 |
| 指导教师  意见 | 签名：  年 月 日 |
| 学院审核  意见 | 盖章：  年 月 日 |
| 学校审核  意见 | 盖章：  年 月 日 |
| 专家组审核意见 | 签名：  年 月 日 |
| 省实施办公室审核意见 | 盖章：  年 月 日 |