学习汇报

|  |  |  |  |
| --- | --- | --- | --- |
|  | 文章名称： | 基于人工智能的信息安全 |  |
| 研究领域： | AI+安全 |
| 汇 报 人： | 杨孟衡 |
| 联系电话： | 15398615549 |

|  |
| --- |
| 一、研究主要目标、研究内容、技术关键、技术路线 |
| 1. 研究目标   近年来，随着计算硬件和算法的快速发展，人工智能(AI)在图像识别、教育、自动驾驶汽车、金融、医疗诊断等广泛领域显示出对人类的显著优势。然而，从最初的数据收集和准备，到训练、推理和最终部署，基于人工智能的系统在整个过程中通常容易受到各种安全威胁。  在这些过程中，主要有两方面的问题。  第一是在基于人工智能的系统中，数据收集和预处理阶段分别容易受到传感器欺骗攻击和缩放攻击，而模型的训练和推理阶段分别容易受到投毒攻击和对抗性攻击。  第二是人工智能、机器学习(ML)和数据挖掘在安全和隐私问题上的潜在应用正在不断扩大。这些技术提供的分析工具和智能行为使得人工智能和学习在具有丰富数据或需要对不断变化的情况做出快速反应的领域的自主实时决策中变得越来越重要。特别是，这些智能技术为涉及通过云计算扩展的大数据分析的安全问题提供了新的解决方案。在安全敏感领域中使用学习方法为安全研究创造了新的前沿，在这些领域中，对手可能试图误导或逃避智能机器。报告主要介绍了人工智能的生命周期以及回顾了人工智能安全问题的挑战和最新研究进展，从而描绘了人工智能安全的总体蓝图。更具体地说，报告以人工智能系统的生命周期为指导，介绍每个阶段出现的安全威胁，阐述人工智能安全与国家安全的关系，然后详细总结相应的对策。最后，还将讨论人工智能安全问题的一些未来挑战和机遇。   1. 研究内容 2. 人工智能生命周期   人工智能生命周期分为五个阶段，分别是数据收集、数据预处理、模型训练、模型推理、系统集成。  在数据收集阶段，安全风险与采集数据的方式密切相关。数据收集方式主要有两种:基于软件的收集和基于硬件的收集。基于硬件的收集方法的一种代表性攻击是传感器欺骗攻击，攻击者通过访问或篡改传感器提供的数据来执行传感器攻击。基于软件的数据收集方法主要是指收集数字数据,其安全风险包括数据偏差、虚假数据和数据泄露。  在数据预处理阶段，目前主要是缩放攻击。缩放攻击一般针对图像域,图像数据在预处理阶段可能被篡改,从而成为潜在的攻击面。具体来说,对于阴险的ISA,攻击者篡改图像并用人与机器之间的(视觉)认知差异,以实现欺骗和逃避攻击,甚至绕过仔细的人工检查。与依赖于模型的对抗性示例攻击不同,ISA仅针对数据预处理步骤。攻击者利用L-P norm来控制目标图像与攻击图像之间的距离,以提高攻击成功率。数据随机化、质量监控、图像重建是打败ISA的主要技术。  在模型训练阶段，致病性攻击通过向模型中注入有毒数据,从而篡改训练模型,从而影响训练数据和训练过程。一般来说,致病性攻击主要是指数据中毒攻击,分为两类,即可用性攻击和完整性攻击。对于可用性攻击,其通常根据模型的梯度信息找到中毒点,或者使用辅助网络自动生成中毒数据，对于任何输入，可用性攻击都会降低模型的整体性能。相比之下,完整性攻击不会影响正常输入的分类,而只会影响攻击者选择的输入。后门攻击和清除标签中毒攻击是代表性的完整性攻击。现有的防御投毒攻击的策略包括数据消毒、鲁棒性训练和认证防 御。在模型推理阶段，逃避攻击通常在模型推理阶段执行,通过制作对抗性示例来降低或干扰模型的预测性能,这些示例通常通过对输入进行较小且语义一致的更改, 但不改变目标模型。这种攻击已经在图像分类、语音识别、 NLP和恶意软件检测中得到了广泛的研究。近年来开发了大量的对抗性样例生成策略,如经典的快速梯度符号法(FGSM)、基于雅可比的显著性图攻击(JSMA)、DeepFool等,主要通过优化搜索或基于梯度的信息实现。相应的,对策也被交互式地设计出来,包括基于模型的策略,如蒸馏、检测器、网络验证,以及基于数据的措施,如对抗性训练[55]、数据随机化和输入重建。  在系统集成阶段，安全问题变得相当复杂。在实际应用场景中,AI应用的系统集成不仅涉及到AI技术本身的安全风险,还涉及到机载系统、网络、软件、硬件的结合点所产生的问题。这些威胁包括AI数据和模型的机密性,代码漏洞,AI偏差等。  人工智能各个生命周期与攻击如图1所示。    图1 人工智能生命周期与攻击   1. 国家安全与人工智能安全   涉及到国家安全，信息安全就是不可避免的。而基于人工智能的信息安全也就越发重要，对于加强综合国力以及维护国家稳定有着不可替代的作用。然而，信息可及性的进步给维护国家安全稳定带来了无数的复杂性。其中社交媒体是最重要的信息来源之一，但是它的发展无疑增加了信息操纵不可控因素，损害了国家安全。为了更好地完成国家安全计划，信息技术可以帮助各国识别潜在威胁，安全地共享信息，并保护其中的机制。人工智能是智能领域之一，它强有力地促进安全信息处理，以避免威胁和网络攻击。经过训练，它可以明智地审查通过社交媒体向公众提供的信息，并协助遏制对国家安全的负面影响。报告主要研究分析以下四个主要的目标：  1)公众可获得的信息；  2)影响国家安全的信息；  3)网络攻击的风险；  4)人工智能对国家安全的信息作用；  报告的主要目标是揭开信息可及性视角的神秘面纱，主要介绍与国家安全相关的信息可及性和不可及性的基本原理。其中为了支持和更进一步说明上文提到的研究内容，在下一小节关键技术将介绍一个针对国家安全的有远见的模型和参考框架。报告总结了基于人工智能工具的分类、算法功能和基于区域的特定领域分析，以突出研究目标。最重要的是，对于许多天真的社交媒体用户来说，这份报告剖析了社交媒体是如何让一些普通人在不知情的情况下如何一步一步成为网络攻击的受害者，随后会危及国家安全。接下来就以国家安全中最突出的信息安全开始分析。  国家安全与信息传播息息相关，安全离不开信息化建设，离不开信息安全。虽然多年来已经开发了各种机制来解决信息处理问题。然而，以社交媒体为基础的信息传播方式，虽然带来了许多好处，如信息传播速度非常快，信息传播成本低。但也表明迫切需要制定新的国家安全机制来应对不断恶化的环境。软件工程研究所(SEI)的计算机应急响应小组(CERT)报告说，自2001年以来，安全事件每年翻一番，增长率不断增高。与这些数字相对应的是，有大量的安全事件从未被报道过。国家安全体系在很大程度上依赖于影响当前和未来变化的信息，而不考虑国家安全利益。国家安全系统在很大程度上依赖于通过 各种反国家渠道传播的信息。  国家安全可以定义为一种国家战略，以确保保护国家的基本和持久需求，同时以人员和社会价值观保护每个公民的基本安全需求。对国家安全利益的剥削操纵了想要的安全领土环境，导致了对国家价值、利益和全球关系的诋毁。目前，社交媒体是一个庞大的平台，数十亿用户每秒匿名发布大量信息，这对维持国家安全的稳定构成了具体威胁。网络上传播虚假信息是一种持续存在的情况，目前毫无防御能力，这给监管违反言论自由原则的信息流动造成了阻碍。 通过各种途径传播的信息没有义务在安全环境内传播。例如，社交媒体上的信息也可能妨碍国家安全。  除此之外，社交媒体信息在信息战中的作用备受争议。它被黑客活动分子和犯罪分子用作网络开发、黑客攻击和敲诈勒索的便捷工具。社交媒体账户和页面对国家安全至关重要，因为数百万用户在特定时间进行交互，为国家的脆弱性和声誉损害提供了巨大的风险。在不付出额外努力的情况下，虚假信息在破坏国家安全措施方面发挥了巨大作用。它已经成为一个党派问题，可能使国家行动计划陷入瘫痪。上网的最大代价是假新闻的传播，也被称为虚假信息或垃圾信息。然而，所有类别的信息(虚假信息、垃圾信息、虚假操纵信息、部分信息)都导致了一种恐惧和不作为的奇特混合。由于其他国家的敌意和破坏国家安全的意图，信息也充斥着外国干涉。为了防止这种威胁的情况，人工智能在数字发展和信息革命对社会的影响中发挥了作用。人工智能支持的社交媒体信息监控工具可以作为社交倾听工具，帮助面对社交信息档案和受众。人工智能参与解释不同尺度的社会数据信息，可以促进调查其中所说的内容，并根据该信息提取罪魁祸首。在处理具有信息动作检查的内容生成时，使用机器人会更智能，对国家安全信息至关重要的人工智能领域的层次表示如图2所示。    图2 对国家安全信息至关重要的人工智能领域的层次表示  最后，概括一下国家安全与人工智能安全相关的分析内容与研究目标。  分析内容主要包括：  1.通过讨论和分析在社交媒体平台上传播的信息对国家安全措施的有效性和作用，以确定信息在国家安全中的作用  2.通过讨论和探索网络空间的竞争、网络攻击、网络盗窃、虚假信息活动以及网络相关因素的潜在风险，以提供一个更广泛和可理解的有远见的推断出实质性的推理，以衡量和防止国家的威胁和攻击  3.人工智能在国家安全救济和协助打击假情报者和各种网络分裂方面的立场和方面，即讨论人工智能对抗国家利益的数字信息战  4.揭示网络空间违法行为的信息圈，简要讨论并突出相互关联的可追溯性和跟踪性  5.设计并提出了一个限制/控制信息的模型，以及可能跟踪和停止针对和破坏国家安全的社交媒体信息的功能算法。  6.通过讨论和探索国家安全方面的挑战和信息缺口，以描绘网络空间违反国家安全的行为  研究目标包括：  1. 通过社交媒体，公众可以获得哪些与国家安全相关的信息?  2. 随着时间的推移，信息如何影响国家安全?  3. 信息驱动媒体造成的网络攻击、网络空间操纵和网络战争的当前和未来风险是什么?  4. 人工智能如何在国家安全中扮演至关重要的角色，完成称职的信息角色?   1. 关键技术 2. 在上一节中我们阐述了研究分析内容，接下来以人工智能生命周期为指导，阐述在各个周期阶段AI受到的攻击与防御技术。   1.0首先是数据收集阶段。数据是什么，数据是AI快速发展的动力，它有许多不同的形式。例如，数据类型包括但不限于:由硬件设备(如传感器)捕获的图像和音频，由计算机系统自动生成的文档和日志，以及由我们的互联网活动产生的那些(如文本、图像、视频、痕迹)。此外，数据收集所涉及的安全问题并不是AI所独有的，它本质上存在于任何需要数据收集的行业。在阅读文献时，我发现大多数现有的数据采集技术满足功能需求，但安全目标却一般达不到要求。虽然对数据收集方法的分类还缺乏共识，一般可分为基于软件的数据采集和基于硬件的数据采集。基于软件的数据采集处于数字世界，而基于硬件的数据采集则是将物理世界中的物理量转化为数字形式的关键点。下面表1总结了与数据采集相关的攻击与防御。  表1 与数据采集相关的攻击与防御   |  |  |  |  | | --- | --- | --- | --- | | 方法 | 安全问题 | 典型场景 | 潜在的防御 | | 基于软件的数据收集 | 数据偏差 | 社交网络  推荐系统 | 检测与过滤  标准化管理 | | 虚假数据 | 物联网  社交网络 | 检测与滤波 | | 数据泄露 | 涵盖了所需数据收集的场景 | 加密或认证 | | 基于硬件的数据收集 | 传感器欺骗攻击 | 物联网 | 输入滤波  传感器增强  基带偏移 |   1.1基于软件的数据采集。互联网用户的日常活动以数字形式产生了大部分数据。数据收集者使用软件程序工具来收集数据(例如，爬虫或“抓取”内容)。基于软件的数据收集需要数据包捕获应用程序、数据包捕获库、操作系统、设备驱动程序和网卡共同工作，以完成数据收集过程。从理论上讲，这个过程中任何一个环节出现问题都会影响数据采集的质量。我们将以在线社交网络为例，讨论基于软件的数据收集方法所带来的安全风险及其相应的防御措施。数据偏 差和虚假数据是社交网络数据收集所面临的具有代表性的安全风险。  1.2基于硬件的数据采集。与硬件相关的数据采集设备包括传感器、硬件探头、移动终端、数据采集生成卡、内联水龙头、网络接口卡、移动终端等。根据硬件的底层设计原理不同，每种数据采集方式的潜在威胁也不同。传感器是应用最广泛的数据收集工具，它们提供了效率和灵活性的优势。我们以传感器数据采集的安全威胁为例，说明基于硬件的数据采集方法的一些典型安全风险。  1.3在数据采集阶段易受到的攻击类型主要有数据偏差，虚假数据，数据泄露，传感器欺骗攻击。  1.3.1数据偏差。AI对训练数据非常敏感。数据源选择和数据准备可能会引入偏差。例如，平台可能受到商业考虑(例如，特定促销)或政治策略的驱动，以“推动”社交网络中的用户行为。此外，社交平台不鼓励第三方收集数据，并对应用程序编程接口(API)施加了许多限制。因此，数据收集者只能收集有限的数据或与平台呈现给普通用户的数据不同的数据。AI的不完全学习偏见引发了各种各样的担忧，比如性别歧视、种族主义等等。例如，亚马逊人力资源部在2014年至2017年期间使用了一款支持人工智能的招聘软件。结果，亚马逊公司雇佣了更多的男性求职者，而降低了女性求职者的简历。Twitter上有人使用PULSE算法，将一张输入模糊的奥巴马图像还原为一张白色五官歪斜的新面孔。虽然不是故意的，但AI偏见破坏了AI的完整性。我们需要改进数据收集标准，开发诊断和减轻偏见的工具。  1.3.2虚假数据。假数据问题并不是AI领域独有的挑战。Wanda等人创新了卷积神经网络的池化功能。此外，他们提出了一种新的动态深度神经网络(DNN)模型算法来检测在线社交网络中的虚假个人资料。Cobb等人讨论了数据收集应用OpendataKit(ODK)在数据收集过程中的安全挑战。他们探讨了IDK数据收集过程中假数据的来源及其防御措施。  1.3.3数据泄露。数据泄露是一个长期存在的问题。Sweeney等人首先发现，只有三个信息字段(地点、性别、出生日期)可以唯一地识别一半的美国人口。值得注意的是，数据泄露不仅是数据收集阶段特有的问题，也可能发生在模型的训练和推理阶段。  1.3.4传感器欺骗攻击。从物理世界产生的数据需要使用相关的传感器元素进行数字化和收集，用于后续的模型训练和推理。传感器无处不在地集成到智能可穿戴设备、自动驾驶车辆和光探测与测距(LIDAR)中，它们是负责数据测量和收集的底层核心组件。攻击者可以利用传感器的物理特性构建恶意样本来欺骗传感器以干扰数据收集。根据目标通道，Shin等人确定了传感器欺骗攻击的三个向量:常规通道、传输通道和侧通道。FooKune等人通过将后门耦合对电路与模拟传感器相结合，进行恶意信号注入，实现了低功率电磁干扰(EMI)攻击。音频信号被麦克风拾取。然后输入信号被放大，EMI通过放大器注入。之后，它们被传输到模数转换器，随后传输到微处理器，最终使电子元件失效。Son等人利用陀螺仪在自身共振频率下的输出会随噪声波动来攻击无人机(UAV)。在陀螺仪的谐振频率下，注入特定噪声会使陀螺仪产生谐振，从而降低精度，干扰UAV的操作。  1.4数据收集可以从硬件安全、软件安全、网络安全三个方面采取数据安全保护策略，缓解安全威胁。此外，保护策略因场景而异。受数据安全策略的启发，建议以下三类作为数据收集保护措施。  1.4.1检测和过滤。Hinnefeld等人研究了AI偏差，并设计了一系列策略(例如，优化预处 理、拒绝选项分类、学习公平表示和对抗性去权重)来检测和减轻AI偏差。为了减轻数据泄露 的威胁，Birnbaum等人提出了一种无监督的离群值检测技术来检测伪造的调查数据，并说明了使用自动数据质量监控的必要性。在硬件数据收集方面，Zhang等人针对不同的攻击锚点进行了基于软件和硬件的防御。他们发现，传感器增强和基带偏移在防御传感器欺骗攻击方面很有用。以麦克风为例，在放大麦克风幅度的同时增加一个低通滤波器，可以抑制20kHz 以上的语音信号，这意味着人类“听不见”的语音命令将被过滤掉。Ignjatovic等人证明了聚合多个数据源进行信任评估的传统迭代滤波算法容易受到串通攻击，因此他们提出了一种收敛性更好、鲁棒性更强的迭代滤波技术来保护传感器网络。我们可以丢弃由缺乏可信度和可信度的收集工具捕获的数据。此外，攻击者可以通过记录和重放用户给出的命令来进行欺骗攻击。虽然过滤是一种方便有效的防御手段，但我们需要警惕引入数据偏差的过滤规则。  1.4.2数据来源与认证。适当的传感器信任机制可以适应禁用从不受信任的设备或未经授权的 设备收集的数据。首先，在通过可信度评估聚合来自传感器节点的数据之前，应该检查传感器 节点的可信度。另一种常用的安全机制是身份验证。例如，Shoukry等人建立了一种物理挑战-响应认证机PyCRA，其中传感器使用物理探针连续主动地感知周围环境。认证机制是通过分析主动响应来检测被操纵的模拟信号以防御恶意传感器攻击来实现的。  1.4.3标准化管理。人为的误用也会影响采集数据的质量，这就需要对相关人员进行管理和培训。因此，我们需要审查数据收集的安全要求(保密性、完整性、身份验证等)，并制定相应的管理程序，以确保数据收集的安全性。此外，建立适当的激励机制可以鼓励数据提供者更诚实地共享其数据，这有利于数据收集的质量。  2.0在数据预处理阶段，主要受到的就是针对图像的缩放攻击。  2.1用于训练模型的图像数据的大小通常是固定的。例如，输入模型的图像通常尺寸为 224×224 或 32×32，由于图像预处理步骤的原因，它比原始图像要小。例如，在数据预处理阶段，需要对图像进行缩放以匹配模型输入大小。图像缩放在保留原始视觉特征并按比例缩放的同时，生成比原始图像在像素方面分辨率更低/更高的新图像。然而，在缩放过程中，攻击者可能会滥用缩放算法来调整像素级信息来制作伪装图像，导致图像缩放前后的视觉语义发生巨大变化。Xiao 等验证了该攻击在多个 基于云的图像服务器(如 Mic rosoft Azur e、阿里云、腾讯和百度图像分类服务)上的有效性。值得注意的是，只要不同的模型使用相同的重新缩放函数来拟合相同的模型输入大小，ISA 就可以对不同的模型不可知。表 2 给出了数据预处理阶段缩放攻击与防御。  表2 缩放攻击与防御   |  |  | | --- | --- | | 攻击策略 | 防御对策 | | 减少目标与攻击图像之间的L-P范数距离 | 随机化  质量监控 | | 减少目标和攻击图像之间的颜色直方图差距 | 图像重建  攻击检测 |   3.0在模型训练阶段，人工智能系统是基于大型策划数据进行训练的。然而，数据质量直接影响训练模型的性能。在这种情况下，攻击者可以毒害训练集来操纵模型的推理行为。从模型和攻击目标的角度来看，投毒攻击可以分为两类:可用性攻击和完整性攻击。  3.1.1可用性攻击被称为拒绝服务攻击，其攻击目标是最大化模型的整体损失，并导致模型性能下降以及错误分类。例如，社交媒体聊天机器人拥有丰富的语料库，并通过与人类的交互进行扩展。当攻击者用一些没有上下文相关性的语句影响聊天机器人时，聊天机器人就不会进行正常的逻辑聊天。  3.1.2完整性攻击是攻击者在不影响模型对干净样本的分类的情况下，通过精心设计有毒数据来实现目标损害的攻击。最具代表性的完整性攻击是后门攻击。后门攻击只会对包含特定(显式甚至不显式)触发器的输入进行错误分类，并且后门仍然可以保留在下游迁移学习任务中。举个后门攻击的例子，在恶意软件检测中，攻击者将包含特定字符串的文件标记为良性数据，并将其放入检测器的训练中。在模型训练和部署后，攻击者只需将特定字符串添加到恶意软件中以逃避检测，因为任何具有特定字符串作为触发器功能的恶意软件都会与良性类相关联。根据攻击行为和分类结果的不同，将投毒攻击分为特定错误攻击和泛型错误攻击。  假设有一个 干净样本 C1，真实标签 *y*true，攻击者构造了一个中毒样本集 C2并将其添加到模型 M 的训练集中， 导致模型M 对 C1进行错误分类，即*M*(*C1*)不等于*y*true。如果M(C)是攻击者所针对的特定类，则为特定错误中毒攻击。然而，如果 M(C)是 *y*true以外的任何类，则它是泛型错误中毒攻击。如图 3.1所示， 实线表示正常情况下的二元分类器。假设在训练集中加入了少量的有毒数据。在这种情况下，决策边界将被移动，从而产生被虚线分隔的分类效果。因此，在正常模型和中毒模型相交形成的封闭区域内的实例将在推理阶段被错误分类。如图 3.2所示，A 类实例将被错误分类为 b 类。   |  |  | | --- | --- | | 图3.1 被错误分类的训练阶段模型 | 图3.2 被错误分类的推理阶段模型 | |  |  |   3.2接下来，将详细介绍各种投毒攻击方法，并且也在表 3 中进行了总结。  3.2.1可用性攻击。可用性攻击被称为拒绝服务攻击。代表性的可用性攻击包括基于梯度的攻击和基于生成式对抗网络 (GAN)的攻击。带有中毒的可用性攻击可以形式化地表示为双级优化问题。内部优化是一个在中毒训练集上的模型训练问题。外部优化是最大化攻击者的目标 A，通常是由内部优化得到的有毒模型上的干净数据集的分类损失函数 L 。其形式化表示如图4所示：    图4 损失函数形式化表示  3.3.1完整性攻击。完整性攻击可以在不影响模型对正常样本分类的情况下完成目标伤害。后门攻击是最具代表性的完整性攻击。后门攻击并不影响在后门模型中被分类的干净数据的结果，但对于包含由攻击者秘密控制的特定触发器的输入，会产生与预期结果的偏差。借壳攻击是一种典型的完整性中毒攻击，通过在干净样本中添加触发器来创建有毒样本，这些样本的标签通常被修改为目标标签。 值得注意的是，触发器，比如它的位置、形状或颜色，都可以在攻击者的任意控制之下。  3.3.2清除标签中毒。Shafahi 等人提出了清除标签中毒攻击，它保留了图像的标签和可视化内容之间的一致性。简而言之，他们通过向训练集中添加有毒数据(标记为基类)来改变模型决策边界，从而导致有毒数据周围的干净目标实例被错误地分类为基类。攻击过程如图 5 所示。    图5 清除标签攻击  3.3.3数据投毒攻击方法与防御  表2数据投毒攻击防御对策   |  |  |  | | --- | --- | --- | | 防御策略 | 优势 | 缺点 | | 数据消毒 | 通用性强，易于实现  在某些场景下容易检测 | 小样本容易过拟合，计算开销高 | | 鲁棒训练 | 模型鲁棒性增强  复杂度低 | 高计算开销 | | 认证防御 | 高可解释性 | 高复杂性 |   表3数据投毒攻击方法   |  |  |  |  | | --- | --- | --- | --- | | 类别 | 攻击策略 | 优势 | 缺点 | | 可用性攻击 | 基于梯度的 | 模型鲁棒性强 | 计算复杂度高 | | GAN-based | 时间开销低  复杂度低  与模型无关 | 泛化程度低 | | 完整性攻击 | 后门攻击 | 不影响正常样本分类  与模型无关  高度隐蔽性  高泛化性 | 局限于图像识别场景，需要重新训练或引入额外的模型 | | 清除标签中毒 | 实现简单 | 计算复杂度高 |   4.0在推理阶段进行的对抗性示例攻击是 AI 系统中研究最多的安全威胁。  4.1图像分类。在图像域，攻击者通过在肉眼无法识别的原始像素上稍微添加扰动来误导模型。对抗性样本的概念最早是由 Szegedy 等人提出的。他们发现，高级神经网络中包含的语义信息(某一特征)分布在整个网络的空间结构中，而不是单个神经元中，神经网络输入和输出之间的映射大多是不连续的。因此，在相同的输入中加入相同量级的干扰， 可以使不同的神经网络产生类似的错误分类。具体来说，神经网络有一定的盲点，可以在输入图像中注入有规律的扰动来欺骗网络。Goodfellow 等人提出的快速梯度符号法 (Fast Gradient Sign Method, FGSM)是图像领域早期具有代表性的对抗样例生成算法。在其他领域，如语音识别、 NLP 和恶意软件检测，数据和分类器结构更复杂，这需要在安装对抗性示例攻击时更加谨慎。  4.1语音识别。自动语音识别 (Automatic Speech Recognition，ASR)是一种使智能设备能够识别和理解人类语音或/并将其转换为文本的技术，需要在提取声学特征之前对原始音频进行滤波和数字化操作。除了MFCC 之外，DFT 和快速傅里叶变换(FFT)也可以提取语音特征。与图像分类相比较领域，ASR 系统的对抗性示例更难制作，因为一些常见的音频处理操作非常容易引入额外的噪声。  4.3自然语言处理。NLP 是用计算机识别人类语言。NLP 的应用范围从输入识别、标签分类到单词、句子的分析理解和处理以及文档的章节。对抗性示例攻击也存在于 NLP 中，尽管在这方面的研究少于图像和音频领域。由于图像像素和文本数据之间的差异，像素级对抗性攻击方法不能直接安装在 NLP 中生成。首先，图像数据(例如，像素值)在数值轮廓中是连续的，但文本数据标签类型是离散的。一般来说，文本数据在输入到 DNN 之前需要进行矢量化。因此，对于文本类型的对抗性示例攻击，大多数研究关注的是阅读理解任务，而不是短文本。  4.4恶意软件检测。恶意软件检测是利用 AI 技术对静态或动态分析提取的软件特征进行分类的过程。静态分析在不执行的情况下提取和分析恶意软件样本的特征。而动态分析则需要执 行并分析其对应的特征。常用的动态分析工具包括沙盒、模拟器等。恶意软件检测中常用的特征有字节序列、操作码、api 和系统调用、网络活动、文件系统、PE 文件等。如图 6 所示，首先，通过特征提取对代表恶意软件的特征序列进行过滤。然后，在由特征序列组成的数据集上训练恶意软件分类器。对抗性示例攻击旨在向特征向量中添加一些功能独立的特征， 以生成恶意软件对抗性示例。    图6 恶意软件检测示例  4.5针对不同的对抗性例子，总结方法优缺点如下表4所示：   |  |  |  |  |  | | --- | --- | --- | --- | --- | | 防御动机 | 防御策略 | 防御方法 | 优势 | 缺点 | | 模型 | 修改网络 | 模型防御蒸馏 | 易于训练，开销低更好泛化，性能好 | 与模型高度相关  高计算复杂度 | | 梯度正则化 | 模型鲁棒性增强 | | 深度压缩网络 | 高鲁棒性 | | 模型改进 | 特征压缩 | 低复杂性 | 降低模型预测性能 | | 额外的网络 | 探测器 | 低复杂性  与模型无关 | 低泛化性  无法增加模型稳健性 | | GAN-based | | 验证网络 | 网络验证 | 高可解释性 | 巨大的计算开销 | | 数据 | 修改  培训过程  或输入数据 | 对抗性训练 | 容易实现  防御能力强 | 难以收敛  高计算开销 | | 数据随机化 | | 数据压缩 |   5.0人工智能技术无处不在。虽然我们已经讨论了 AI 本身的威胁和对策，但当 AI 集成到现实世界的应用程序中时，安全问题似乎更加复杂。对于不同的应用场景，安全问题是不同的，我们应该从全局的角度来看待 AI 的安全。本节将探讨实际集成阶段的几个安全风险。  5.1人工智能机密性，AI的机密性包括数据机密性和模型机密性。AI的机密性通常与模型隐私明确相关，尽管它也可能(间接)导致安全问题，如模型反演和模型提取。模型反演是指基于输入和输出之间的映射关系，对模型进行逆分析，以获得私有数据。另一方面，模型提取通常被理解为通过API执行可接受数量的查询，并观察输出结果(概率或标签)来推断模型参数或提取与目标模型密切匹配的近似模型。对于这两种类型的隐私问题，通常使用dp差分隐私、同态加密或模型水印来减轻隐私风险。目前的AI安全和隐私问题似乎是分开解决的。尽管具有挑战性，但值得考虑系统地、并发地解决这些问题，以确保数据和模型隐私，同时保持AI系统的安全性。由Google研究团队提出的联邦学习是一种新的分布式机器学习技术，已成为AI的另一个重要新分支。联邦学习聚合了在每个客户端持有的本地化数据上训练的本地模型，以更新全局模型。联邦学习主要缓解了隐私问题，但它缺乏对本地数据的审计和对参与者行为的控制，这很可能会引入安全问题。联邦学习的客户端、中央服务器和通信渠道很容易成为攻击者的目标。最典型的是投毒攻击，可以通过数据投毒或/和模型投毒来实现，用户端/服务器端GAN攻击，以及联邦学习中的隐私问题。由于缺乏对数据的访问权限和对客户端的控制有限，设计针对安全攻击的对策比集中训练更具挑战性。  5.2代码漏洞保护。当前的AI系统技术(如深度学习)是建立在框架上的(如Tensorflow、Caffe和Torch)。这些框架依赖于各种基础库和第三方组件，这些组件极大地促进了AI技术的发展。然而，它们并不是设计得完美无缺，存在漏洞。更核心的是，最近的一项研究表明，未经审计的第三方代码片段——损失计算的代码响应——可以在深度学习模型中植入后门。虽然代码漏洞是模型实现过程的一部分，但它们也是AI系统安全部署的关键部分。  （2）在上一小节中我们讨论了人工智能生命周期中可能遭遇的攻击与防御对策，接下来以国家安全与信息安全的关系为指导阐述信息安全关键技术。其中，国家安全与信息安全的关系以社交媒体与国家安全为例如图7所示。    图7网络空间的信息周期及其与社交媒体的融合  图7说明了网络空间中的信息周期，以及社交媒体作为其中一个组成部分是如何相关的。社交网络应用程序和社交系统信息学是一个协调的纽带， 以促进数据合成、数据操作和第三方数据共享。社会系统信息学管理与网络应用程序密切相关的数据源和频率。它更像是一个高度连接的特征，其中一个信息集触发其他信息集，循环继续，直到一个合适的受众被定位和告知。而社交网络应用程序方便并提供用户的帐户信息，他们的共享信息以及搜索和兴趣历史。 最重要的是，音频/视频和文本检索包含可用于国家 安全的可信信息。这组信息还有助于操纵用户进行欺诈、勒索，有时还可以利用一个人来损害国家利益。  如图7所示，另外两个主要因素是社交网络数据库和数据馈送与交换。有组织形式的信息创建了数据库。社交网络数据库提供存在于数百万用户之间的完整链接集。例如，我们国家所有居民的数据库应该具有相同的数据链接，并且可以根据年龄、种族和族裔群体等因素在几秒钟内广播所需的信息。任何意图操纵的信息都可能在这种联系的基础上传播，从而危及国家安全和国家行动计划。图7还展示了社交网络数据库对未来数据预测、活动 跟踪、市场趋势和用户响应任何国家威胁的响应率的贡献。数据馈送功能扩展了国家妥协。即使一个国家的居民居住在该国以外的地方，数据馈送仍然会通过基于数据链接的手机发送，从而引发愤怒。这一因素也可以通过未经同意的第三方数据共享来实现。社交媒体的力量已经发现了社会构建的信息炸弹，这些炸弹可以在几秒钟内成为国家安全的目标灾难，而无需付出任何代价。  通过传统方式和网络方式传播信息具有不同的传播强度和后果。网络战争使普通人的信息共享能力超越了控制能力,受众和平台的变化现在成为国家反安全机构的首选武器。为每个使用者提供信息共享能力，应通过定义好的过滤器，以保护任何负面和不可预见的情况。  我们也需要理解国家安全框架需要人工智能。目前在线社交网络一直呈现出爆炸性的年增长。在成为使用最广泛、适应性最强的信息源后，它极大地迎合了国家利益。定向广告和病毒式营销被引入，在没有事先检查和审查的情况下利用社会信息。由于媒体平台上的再分享功能，虚假信息往往传播得更快。 要删除虚假信息，需要将其从所有已经共享该信息的相关人员中删除，者往往带来非常大的安全成本。移除威胁性信息是一项看似乏味的任务，但消除其对 人们的影响更具威胁性和挑战性。人们的日常生活越来越依赖于在线获取信息及其相关服务。通过网络空间对社交媒体用户的操纵既高效又方便。有效区分 授权和未经授权的信息源可以极大地帮助在第一时间阻止错误信息。   1. 技术路线   以国家安全与信息安全关系为例阐述技术路线，通过图8所示中的检查技术路线，可以使得信息变得可停止和可问责。    图8 技术路线模型  在图8中，提出了一个简单的模型，它可以明智地采取行动，限制信息的传播，最大限度限制国家的妥协，根据该模型，信息来源于两种不同的源头：  1.真实来源，如批准用户、国家组织和政府控制的媒体渠道  2.非真实来源，如数百万社交媒体用户和私营信息广播集团  需要保证来源于非真实来源的信息源的服从于AI框架。  经过AI框架过滤和检查后，信息要么被批准要么被抛弃。模型中提出的人工智能框架可以解决虚假信息传播、反国家用户、稳定威胁、信息过滤等问题。该模型封装了每个信息上下文，即来源验证，内容和文本分析，图像和语音分类以及知识构建，所有这些都由人工智能机制提供支持。此外，它还可以帮助定位信息来源。在不真实的源可以执行某种操作之前，应该始终对它们进行检查。 |

|  |
| --- |
| 二、项目的创新性(理论创新、技术创新等方面，根据学习材料的具体内容描述)600字及以上 |
| 本项目/文章针对基于人工智能的信息安全难题，提出了不同的解决方案，研发了算法，实现了信息安全检查模型，有效解决人工智能安全与信息安全难题。具体研究意义与创新点如下：   1. 现有的信息安全针对国家安全的调查既不全面，也不是基于系统文献综述进行阐述的。此外，据我所知，现有的信息安全并没有涵盖社交媒体信息及其对国家安全的威胁的所有方面，因此存在一定的不确定性 2. 现有的信息安全领域针对国家安全的调查不是最新的，也不包括人工智能在信息安全和国家安全措施领域的最新工作，很难找到前沿基于人工智能的信息安全文献 3. 在早期文献中没有找到与基于人工智能的信息安全相关的内容，特别是对人工智能方法的特定应用分析，以及对人工智能生命周期某个阶段研究分析的分析报告 4. 通过功能算法设计和支持信息身份验证和身份验证模型，并通过一系列步骤利用AI 框架进行工具辅助 5. 通过探讨社交媒体对国家安全的影响，以当前二十一世纪新时代下的互联网时代为背景，阐述了信息安全如何影响国家安全，信息安全如何影响其他安全，如何对我们日常活动造成不利的影响 6. 报告详细介绍了 AI 系统在数据收集、模型训练、模型推理和集成应用等方面的安全研究进展，并对相关防御技术进行了综述 7. 以综述的方式详细介绍了人工智能生命周期各个阶段易于受到的攻击，以及对应的防御对策，之后针对不同阶段攻击提出了不同的防御方法，说明了具体的防御措施，最后以国家安全为例设计了安全模型 8. 本报告主要意义是研究分析人工智能生命周期与基于人工智能的信息安全工具，比较分析基于人工智能的信息安全对国家安全的影响，以启发与网络空间侵犯、社交媒体操纵者和虚假信息运动相关的威胁因素及其强化效果、措施和可能的预防措施 |

附件：知识树

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