

**硕士学位论文开题报告**

**基于对抗样本的版权算法研究**

姓 名： 王 杰

学 号： 2111915024

导 师： 丁 烨

学科专业（代码）： 081200

研究方向： 计算机网络与信息安全

东莞理工学院研究生处

2020年9月版

**硕士研究生学位论文开题报告**

**撰写要求**

一、只有学籍状态为注册或悬置的研究生才允许开题。但学籍状态为悬置的研究生只有在完成注册手续之后，开题报告及其评审结果才能被认可。

二、开题报告主要内容

1. 选题的依据包含选题来源及意义、国内外研究动态（不少于4000字）和主要参考文献；

2. 研究方案及工作基础，包含研究目标及拟解决关键问题、拟采用的研究方法、特色与创新性和现有工作基础；

3. 论文工作计划及预期成果；

4. 作者承诺及导师意见；

三、开题报告会由开题报告评审小组（5人及以上的单数）进行集体评议，其中应包含2位及以上具有丰富学术学位硕士培养经验的外校导师。评审小组应根据开题报告的内容和表现情况提出具体意见和建议并给出结论，开题报告结论分为通过和不通过两种。

四、开题报告后，研究生应根据评审小组的意见和建议，对选题方案进行修正、补充和提高，完善《硕士研究生学位论文开题报告》，按规定的程序报批备案后，方可进入论文写作阶段。

五、开题报告后，若学位论文课题有重大变动，应重新作开题报告，并按程序重新报批。

六、研究生开题报告考核等级为不通过，该生应在3个月内重新开题。

七、《硕士研究生学位论文开题报告》填写内容必须属实，双面打印，一式两份，由研究生本人和学位点存档备查。

八、学位论文（设计）开题报告打印格式：

1．打印用纸：A4。

2．页面设计：上下左右边距分别为3cm、2.5cm、2.5cm和2.5cm。正文字体、字号：宋体、小4号。图题字体、字号：宋体、小5号。正文行间距22磅，图题单倍行距。页码设置奇偶不同，居右排版。

**一、摘要**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 论文题目 | 中文 | 基于对抗样本的版权算法研究 | | | |
| 英文 | Research on Copyright Algorithm Based on Adversarial Examples | | | |
| 导师组 | | 姓名 | 职称 | 工作单位 | 研究方向 |
| 导师 | | 丁烨 | 副教授 | 东莞理工学院 | 计算机网络与信息安全 |
| 指导小组 | |  |  |  |  |
|  |  |  |  |
|  |  |  |  |
| 学位论文所属类型 | | □基础研究 □应用研究 □综合研究 □其他 | | | |
| 课题来源 | | ⑴ 国家重点研发计划□ ⑵国家社科规划、基金项目□  ⑶教育部人文、社会科学研究项目□ ⑷国家自然科学基金项目☑  ⑸中央、国家各部门项目□ ⑹省（自治区、直辖市）项目□  ⑺国际合作研究项目□ ⑻企、事业单位委托项目□  ⑼外资项目□ ⑽学校自选项目□  ⑾国防军工项目□ ⑿非立项□ ⒀其他□ | | | |
| 论文选题项目是否涉密 | | □ 是，并已办理《东莞理工学院学位论文定密审批表》审批手续，复印件已附于本报告末页。  ☑ 否，并承诺学位论文将不涉及国家秘密和其他不宜公开的内容。 | | | |
| 摘 要（不超过800字） | | 数字水印可以将某些特定的信息嵌入到数字媒体中，以实现版权保护。常见的水印技术主要分为空间域和变换域。然而，这些方法大多都存在鲁棒性差、感知度高、容量大的问题，并且在某些机器学习算法下易被检测和掩盖。 为了克服上述问题，我们提出了一种全新的数字水印系统。该系统将目标检测模型与对抗攻击相结合生成特殊的数字水印（即对抗样本）并将其嵌入到源图像中。根据生成的对抗样本的原理以及我们在现实生活数据集上的实验，本系统具有以下优点：1）水印肉眼察觉不到；2）水印在每个输出图像上都是唯一且可追溯的；3）水印具有强大的防御干扰能力。 | | | |
| 关键词 | 中文 | 数字水印 目标检测 对抗样本 | | | |
| 英文 | digital watermark/object detection/adversarial example | | | |

**二、选题依据**

|  |
| --- |
| 1、选题来源及研究意义  近年来，数字内容的版权保护成为了重要的研究领域，尤其是在中美贸易战之后，知识产权成为了其中最主要的需求之一。保护数字内容版权的方法有很多，数字水印[1-2]是主要方法之一。数字水印旨在将特定信息嵌入到原始媒体中，包括图像，音频和视频。根据水印的目的，水印可以是可见的也可以是隐藏的。如果水印旨在显示版权信息，例如，作者希望其他人共享其原始作品，但必须指定作者信息，则通常会使用可见的水印。相反，如果作者禁止共享原始作品，则可以将隐藏的水印应用于源跟踪：将追踪和指责共享原始副本的任何人。对于水印的后一种形式，隐写术是一种在原始信号中隐藏信息的有效方法，该信息可以嵌入到像素或者量化到变换域中。数字水印正是信息隐藏技术的一个分支隐写方法，一般通过以下指标[3]来评估强大的数字水印算法：  1）鲁棒性：指示嵌入的水印能够防御某些篡改攻击的能力。例如，如果旋转，调整大小或裁剪嵌入的图像，则图像内部的水印仍应是可识别的或部分可识别的，也就是高鲁棒性。较好的水印技术一般具有较强的鲁棒性。  2）感知度：指示嵌入的水印的可见度。对于隐写术，通常嵌入的水印应不可区分或几乎不可见，也就是低可感知性。较好的水印技术一般具有较低的感知度。  3）容量：指示可以嵌入的水印的大小。一般而言，较高的容量将导致较差的鲁棒性和较高的可感知性。较好的水印技术一般具有较小的容量。  然而现有的数字水印技术，在鲁棒性和感知度上存在着一定的不可兼容的问题，并且在一些特定的机器学习算法中也容易被检测和掩盖，容易造成版权保护方法的失效。因此，本课题寻找一个同时能够对抗机器学习算法的检测，也能调节鲁棒性和感知度之间存在的矛盾的新型水印技术。为此本课题引入了对抗样本[4]概念，一个不可感知却足以误导机器学习模型分类错误的微小扰动。参考其产生的原理和特性，将其视为一种特殊的信息隐藏技术，并提出了一个基于对抗样本结合目标检测的新型水印系统，以实现版权保护，十分具有研究意义。  2、文献综述  2.1 数字水印  现有的大多数数字水印方法是将特定的信息（例如一个图像标签，一段句子或一串可识别的符号）嵌入到原始媒体中。图1为一个常见的数字水印系统，主要包括三个部分：水印嵌入、攻击测试、水印提取。其完整的一个流程为，给定一个原始信号S，经过嵌入函数E将一些指定的信息嵌入到S中得到SA，之后经过攻击函数（例如几何操作、噪声、滤波等攻击）得到SEA，最后经过检测恢复函数D和R，输出最后的提取结果。  信号  S  嵌入函数  E  攻击函数  A  检测恢复函数D R  S  SA  SEA  结果  图1水印系统  数字水印根据使用的载体不同，主要有图像、音频、视频、文本等多媒体数据流。根据使用的功能分为可见水印以及不可见水印。本课题主要针对于图像的不可见水印的研究。  传统的数字水印技术主要分为两类：空间域和变换域。  在空间域中实现的典型方法是最低有效位（Least Significant Bit，LSB）。这种方法的主要原理是利用在像素强度变化很小的情况下去修改图像的重要部分。例如，将图像切分成相等大小的子块，然后把特定的水印嵌入到子块中，常见的水印嵌入方法是直接修改图像的像素值。这类水印技术的实现方法相对比较简单而且计算成本很低，但是这种方法生成的水印具有较差的鲁棒性而且感知度相对较高[5]（肉眼可见）。对于常见的攻击，包括例如缩放、旋转、剪切等几何操作[6]、JEPG压缩操作、滤波和噪声等攻击就无法正常提取到水印了。之后Bamatraf等人[7]提出了第三和第四最低有效位插入水印技术，它比传统的LSB方法具有更强的鲁棒性和更高的水印质量，并且在一定程度改善了LSB方法的不足。而其他的方法包括中间有效位（Intermediate Significant Bits，ISB）或拼凑算法（Patchwork），以及扩频和其相关的算法。虽然在空间域上的水印技术在不断更新迭代，但这种在空间域上生成水印的技术存在着一定的缺陷。它的生成原理都是基于图像像素的直接修改，因此生成的水印普遍存在鲁棒性较差和感知度较高的冲突，并且这类水印只有在不暴露于任何噪声或者不遭受任何攻击的情况下才能表现出其性能。  与空间域的水印算法不同，在变换域的水印技术主要是通过改变原始图像的某些特定属性（比如颜色，纹理，频率）来实现。具体的实现如图2所示，使用预定义的转换手段来转换图像，以便在频域中表示图像。然后，再通过使用不同的变换来更改原始图像的变换域系数来嵌入水印。这种变换域的水印算法在一定的程度上它极大地提高了水印的鲁棒性同时也降低了水印的可见度。  宿主图像  应用预定于的变换  例如，DCT/DFT/DWT/SVD  水印嵌入  水印图像  通信通道  水印提取  对宿主图像应用反变换  水印图像  攻击  图2水印的嵌入和提取  变换域主要包括离散余弦变换域（Discrete Cosine Transformation，DCT），离散小波变换（Discrete Wavelet Transformation，DWT）域、离散傅立叶变换域（Discrete Fourier Transformation，DFT）和奇异值分解（Singular Value Decomposition，SVD）等以及它们之间的混合变换域。其中，DCT变换能量集中，实现简单，为图像压缩标准使用，利于提升不可感知性。1995年，Cox等人[1-2]最先将数字水印嵌入在原始图像的DCT域中，这是最早的在变换域中实现水印的嵌入的方法。其主要思想为：首先将原始图像通过DCT变换转换成频域图像，并且得到其DCT系数S序列，从S中选择n个重要的频率段组成S1，将水印信息嵌入到S1中。后面的研究[8-11]都是基于文献[1, 2]的展开。例如基于图像分割和离散余弦变换（DCT）的水印算法，首先是使用期望最大化（Expectation Maximum，EM）算法对图像进行分割。对于每个片段，将图像片段细分为大小为8乘8（64像素）的像素块，并对其之字形进行重新排序，然后计算该块的DCT，之后将实数的伪随机序列嵌入每个图像段的DCT域中。这种水印方案对常见的水印攻击（例如，旋转，剪切操作）具有较高的鲁棒性[8]。之后，伴随着神经网络的出现，提出了一种基于离散余弦变换（DCT）和神经网络的鲁棒数字图像水印技术。该神经网络采用的是全连接神经网络（FCNN），当时FCNN被用来模拟原始图像的感知和视觉特征，而原始图像的感知特征可以查找DCT系数的最高的可更改阈值。这个最高的可变阈值是将水印嵌入原始图像的DCT系数中重要参考。由于水印是二进制图像，图像的像素值在图像的DCT系数中作为0和1插入[9]。为了获得更好的性能，Rao等人[12]将水印嵌入到宿主图像的DWT系数的DCT域中，与单独使用DCT相比，具有更好的性能。SVD变换具有唯一确定性、改动幅度小、空间复杂度低等优势，利于提升鲁棒性。基于SVD混合域嵌入算法也受到不少研究者青睐。在参考文献[109]中，将DCT，DWT和SVD组合在一起以获得更好的鲁棒性和不可感知性。事实证明，这种混合方案对于JPEG压缩，过滤，椒盐噪声，旋转，裁剪以及缩放和平移操作都具有鲁棒性。  根据水印的评估指标，上述两种传统水印技术虽然具有较小的容量，但可能也会导致鲁棒性降低和感知力升高。此外，可识别的信息也可能会引起攻击者的兴趣，因为黑客的攻击结果易于理解，所以更容易构成针对性攻击。并且现有的大多数水印方法都存在鲁棒性和感知度的矛盾，并且在某些特定的攻击下（如深度学习算法模型）易被检测和掩盖。为了克服上述传统水印鲁棒性较低和感知度较高导致两者不能很好兼容的不足，并且考虑到传统水印技术在现有的机器学习算法上的一些弊端，我们寻求通过对抗机器学习的方法来设计一个新颖的水印系统。  2.2 对抗攻击  随着深度学习在许多任务上（包括图像识别，语音识别，目标检测，实例分割等）取得了不俗的表现性能，越来越多的工作研究都在利用深度学习进行问题求解。然而这样强大的学习模式，在面对对抗样本的时候也是脆弱的，容易遭受到对抗攻击而严重影响性能，研究对抗攻击不仅能够了解模型的不足，提高模型的鲁棒性，也能防患于未然，避免损失严重。  对抗攻击的常见分类（如图3所示）：   1. 白盒攻击（White-box attack），攻击者完全清楚模型的架构、细节以及训练集。 2. 黑盒攻击（Black-box attack），攻击者不了解，或者了解很少模型以及训练的信息。 3. 目标攻击（Targeted attack），让分类网络错分类到一个指定的类别上。 4. 非目标攻击（Non-target attack），让分类网络错分类到任意错误类别，而不需错分类到指定的类别。   白盒攻击  对抗攻击  决策边界近似近似  基于梯度的优化  梯度近似  黑盒攻击  约束优化  遗传算法  基于搜索  目标攻击  非目标攻击  图3 对抗攻击的常见分类  对抗攻击的发展过程主要有以下两个方向：  （1）一个是不断利用新的算法进行对抗攻击。  2013年Szegedy等人[4]是最早提出对抗样本问题的人，他们发现深度学习网络可以通过添加几乎无法察觉的扰动来轻松对图像进行错误分类，这是对抗攻击的最开始的理论基础。并且在文献[4]中，他们提出了通过BFGS来求解最小的损失函数添加项，让网络分类错误，从而将复杂的限制优化问题转化为凸优化问题求解。2014年，在研究深度网络本质的工作[13]中，Goodfellow等人指出对抗样本出现的原因并非是深层神经网络的高度非线性和过拟合，即使是线性模型也存在对抗样本。同时也提出了一种简单的基于梯度信息优化的单步攻击方法来寻找对抗样本，称为快速梯度符号方法（Fast Gradient Signal Method，FGSM）。这是白盒攻击下的最早的两个主要方向。在此之后也吸引着更多的关注在对抗攻击上的工作，包括DeepFool [14]，Carlini and Wagner攻击(C&W)[15]和Project Gradient Descent (PGD)[16]等等，他们的工作都揭示了广泛使用的深度学习模型中的安全性问题。  为了使深度模型更加健壮，许多研究人员提出了针对对抗性样本的防御方法。在防御中，防御性蒸馏[17]和对抗训练是针对FGSM的两种有效防御方法，但易受迭代攻击的影响[18]。后来研究人员也将对抗学习划分了两条路径：攻击（创建对抗样本）和防御（保护模型）。一种尝试破坏模型的防御，而另一种尝试保护模型。除了这两种策略，还有一些进行对抗样本的检测。Metzen等人[19]提出了一个进行对抗训练的“检测器”网络，它可以明确地检测对抗样本。Feinman等人[20]是另一方面的隐式方法，其中它通过查看可在dropout神经网络中使用的贝叶斯不确定性估计，以及通过对模型学习的深度特征的子空间执行密度估计，来研究对抗性样本的模型置信度。由于深层网络的性质，该模型容易受到对抗样本的影响。对抗攻击和防御的相互博弈，互相促进着双方的发展，越来越多的对抗攻击算法也在不断的涌现[21-22]。  （2）另外一个是不断寻找新的应用场景。  在计算机视觉领域中，包括了对分类器的攻击以及对语义分割和目标检测任务的攻击或者寻找其他领域的对抗攻击（本课题主要针对图像）。  作为计算机视觉中最为经典的工作，图像分类任务上的对抗攻击思路都是通过按顺序或者随机一个一个修改图像的像素。比如Jacobian-based Saliency Map Attack (JSMA)通过计算神经网络前向传播过程中的导数来生成对抗样本[23]。One Pixel Attack 利用差分进化算法[24]，不断迭代变异寻找到足以只修改一张图片中一个像素点来对网络进行攻击，这是一个典型的黑盒攻击方法，它完全不需要知道网络架构、参数等一些细节信息。而语义分割和目标检测任务的对抗攻击要比分类任务难很多，并且目标检测任务的对抗攻击效果更为丰富（基于分类，包含更多的信息，比如坐标信息，检测框等）。  与先前的工作不同，通常将目标检测模型的对抗样本假定为物理样本，而不是不可感知的样本。研究者认为可以创建误导模型的物理对象的攻击者是更强大的对手。在这种情况下，Kurakin等人[25]首先提出了一种威胁模型，该威胁模型可以通过在纸上打印数字对抗样本来创建对抗对象，最终使检测器蒙蔽。Athalye等人[26]改进了Kurakin等人的工作通过创建对视角变化具有抵抗力的对抗对象。Eykholt等人[27]提出了一种对合成和物理图像转换建模的攻击算法。在之后的工作中，Eykholt等人[28]将其先前的算法扩展到更广泛的目标检测模型。  现有的目标检测算法主要分为两个大的类别：基于候选区域分类和回归的模型和直接进行分类回归的模型。前者典型的有以R-CNN[29]，Faster-RCNN[30], Mask-RCNN[31]为代表算法，这类方法是一种分两步执行的程序，检测时长较长精度高。它首先划分出候选区域，然后对候选区域进行分类和回归，最后输出候选区域的类别和位置信息。而后者是单步执行的程序，检测时长短但精度较前者低。这类方法的典型代表算法有YOLO[32]和SSD[33]，它们通过回归的方法直接预测边框的坐标信息。针对于第一类目标检测算法，Xie等人[34]首先提出了一种称为DAG的实现方法，他们为候选区域分配了一个对抗标签，然后迭代更新梯度以对该区域进行错误分类。Chen等人[35]和Li等人[36]随后提出了类似的方法。但上面的方法都是基于目标检测算法的第一类方法，并且攻击时长较长。于是Wei等人[37]提出了一个统一的攻击模型可以同时成功攻击上述两类目标检测算法。但上述在目标检测模型上的对抗攻击研究中，存在攻击时长和攻击效果差（降低了图像的保真度）相互矛盾的问题。  最近的工作中，Brown等人和Karmon等人分别提出了Adversarial Patch和LaVAN，这两种本地化和可见性对抗样本，从根本上改变了图像的一小部分，因此可以将其应用于任何图像并对其分类错误。将这种干扰（对抗样本）视为攻击者嵌入图像中的水印是很自然的。于是，之后Quiring等人[38]将这种干扰与数字水印结合起来进行研究，发现在深度机器学习中很容易检测到传统嵌入的水印信息。  综上所述，针对于水印的第二种形式：隐藏不可见且可追溯，为了保证隐藏的水印不仅能够追溯泄漏源和不影响人们肉眼对图片的视觉感觉，而且还能有效地对抗机器学习模型的检测和修改。本课题设计思路是利用对抗样本的生成原理和特性设计一个全新的水印系统来对抗机器学习模型，从而使其失效，并且还能在一定的程度上解决传统水印在鲁棒性和感知度上存在的不可兼容矛盾。本课题提出的新型水印系统是一个全新的水印系统，主要贡献主要有以下几点：   1. 本课题提出了一种全新的基于对抗样本的水印系统，有别于传统的数字水印技术，极大提高了对抗机器学习检测和修改的能力。 2. 提出了改进的对抗攻击算法，并且实现了对单步目标检测模型的攻击。 3. 主要参考文献 4. Cox I, Kilian J. Secure spread spectrum watermarking for images, audio and video[C]// In Proceedings of IEEE International Conference on Image Processing. IEEE,1996: 243-246. 5. Cox I, Kilian J, Leighton T. Secure spread spectrum watermarking for multimedia[J]. IEEE Transaction on Image Processing. 1997, 6(12): 1673-1687. 6. Miglani E, Gupta S. Digital Watermarking Methodologies-A Survey[J]. International Journal of Advanced Research in Computer Science and Software Engineering Research Paper Available online at: www. ijarcsse. com, 2014, 4(5). 7. Szegedy C, Bruna J, Sutskever I, et al. Intriguing properties of neural networks[C]// arXiv preprint arXiv. 2013: 1312.6199. 8. 李华, 朱光喜, 朱耀庭. 一种基于人眼视觉感知模型的数字水印隐藏方法[J]. 电子学报, 2000(10): 111-113+110. 9. 高珍, 台莉春, 张志浩. 数字水印对抗几何攻击方法的研究[J]. 计算机工程, 2006(05): 135-137. 10. Bamatraf A, Ibrahim R, Salleh M. Digital watermarking algorithm using LSB[C]// 2010 International Conference on Computer Applications and Industrial Electronics. IEEE, 2010: 155-159. 11. Badran E, Ghobashy A, El-Shennawy K. DCT-based digital image watermarking VIA image segmentation Techniques[C]// 2006 ITI 4th International Conference on Information & Communications Technology. IEEE, 2006: 1-1. 12. Nilchi A, Taheri A. A new robust digital image watermarking technique based on the Discrete Cosine Transform and Neural Network[C]// 2008 International Symposium on Biometrics and Security Technologies. IEEE, 2008: 1-7. 13. 钮心忻, 杨义先. 基于小波变换的数字水印隐藏与检测算法[J]. 计算机学报, 2000(01): 21-27. 14. 刘瑞祯, 谭铁牛. 基于奇异值分解的数字图像水印方法[J]. 电子学报, 2001(02): 168-171. 15. Rao V, Shekhawat R, Srivastava V. A DWT-DCT-SVD based digital image watermarking scheme using particle swarm optimization[C]// 2012 IEEE Students' Conference on Electrical, Electronics and Computer Science. IEEE, 2012: 1-4. 16. Goodfellow I, Shlens J, Szegedy C. Explaining and harnessing adversarial examples[J]. arXiv preprint arXiv, 2014: 1412.6572. 17. Moosavi-Dezfooli S M, Fawzi A, Frossard P. Deepfool: a simple and accurate method to fool deep neural networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2016: 2574-2582. 18. Carlini N, Wagner D. Towards evaluating the robustness of neural networks[C]// 2017 IEEE Symposium on Security and Privacy (SP). IEEE, 2017: 39-57. 19. Madry A, Makelov A, Schmidt L, et al. Towards deep learning models resistant to adversarial attacks[J]. arXiv preprint arXiv, 2017: 1706.06083. 20. 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 (SP). IEEE, 2016: 582-597. 21. Carlini N, Wagner D. Adversarial examples are not easily detected: Bypassing ten detection methods[C]// Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security. IEEE, 2017: 3-14. 22. Metzen J, Genewein T, Fischer V, et al. On detecting adversarial perturbations[J]. arxiv preprint arxiv, 2017:1702.04267. 23. Feinman R, Curtin R, Shintre S, Gardner A B. Detecting adversarial samples from artifacts. arXiv preprint arXiv, 2017:1703.00410, 2017. 24. 张思思, 左信, 刘建伟. 深度学习中的对抗样本问题[J]. 计算机学报, 2019, 42(08): 1886-1904. 25. 纪守领, 李进锋, 杜天宇, 李博. 机器学习模型可解释性方法、应用与安全研究综述[J]. 计算机研究与发展, 2019, 56(10): 2071-2096. 26. Papernot N. et al. The limitations of deep learning in adversarial settings[C]// 2016 IEEE European Symposium on Security and Privacy (SP). IEEE, 2016: 372-387. 27. Su, Jiawei, et al. One pixel attack for fooling deep neural networks[J]. IEEE Transactions on Evolutionary Computation 2019, 23(5): 828-841. 28. Kurakin A, Goodfellow I, Bengio S. Adversarial examples in the physical world. arXiv preprint arXiv, 2016: 1607.02533. 29. Athalye A, et al. Synthesizing robust adversarial examples[C]// International Conference on Machine Learning, IEEE, 2018: 284-293. 30. Evtimov I, Eykholt K, Fernandes E, et al. Robust physical-world attacks on machine learning models[J]. arXiv preprint arXiv: 2017: 1707.08945. 31. Song D, Eykholt K, Evtimov I, et al. Physical adversarial examples for object detectors[C]// 12th Workshop on Offensive Technologies, 2018. 32. Girshick R, Donahue J, Darrell T, Malik J. Region-based convolutional networks for accurate object detection and segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015: 38(1), 142-158. 33. Ren S, He K, Girshick R B, et al. Faster R-CNN Towards real-time object detection with region proposal networks[C]// 2015 Conference and Workshop on Neural Information Processing Systems, IEEE, 2015: 91-99 34. He K, Gkioxari G, Dollar P, et al. Mask R-CNN[C]// 2017 IEEE International Conference on Computer Vision, IEEE, 2017: 2980-2988 35. Redmon J, Divvala S, Girshick R, et al. You only look once: unified, real time object detection[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2016: 379-387. 36. Liu W, Anguelov D, Erhan D, et al. SSD: single shot multibox detector(C)// European Conference on Computer Vision, IEEE, 2016: 21-37 37. Xie C, Wang J, Zhang Z, Zhou Y, et al. Xie C. et al. Adversarial examples for semantic segmentation and object detection[C]// Proceedings of the IEEE International Conference on Computer Vision. IEEE, 2017: 1369-1378. 38. Chen S. T. et al. Shapeshifter: Robust physical adversarial attack on faster r-cnn object detector[C]//Joint European Conference on Machine Learning and Knowledge Discovery in Databases. IEEE, 2018: 52-68. 39. Li Y. et al. Robust adversarial perturbation on deep proposal-based models[C]// arXiv preprint arXiv, 2018: 1809.05962. 40. Wei X. et al. Transferable adversarial attacks for image and video object detection [C]// arXiv preprint arXiv, 20181811.12641. 41. Quiring E, Rieck K. Adversarial machine learning against digital watermarking[C]// 2018 26th European Signal Processing Conference (EUSIPCO). IEEE, 2018: 519-523. |

**三、研究方案及工作基础**

|  |
| --- |
| 1、研究目标、研究内容及拟解决的关键问题  1.1研究目标  本课题的主要目标是：在两类目标检测模型上进行对抗攻击，将产生对抗样本的过程视为一个特殊的信息隐藏的手段，利用这种特殊的信息隐藏手段设计水印系统。本系统具有以下优势：1）能对抗机器学习算法的检测；2）提取水印不需要原图对比；3）水印生成速度快，保真度高，几乎不可见。  1.2研究内容  1）在目标检测算法上进行对抗攻击生成对抗样本。  2）将对抗样本加入目标检测模型进行对抗训练。  3）利用生成的水印（即对抗样本）编码。  4）将对抗训练后的目标检测模型改造成解码系统。  1.3拟解决的关键问题  1）如何在目标检测算法进行快速对抗攻击，且能对抗机器学习算法的同时保持图像的保真度和降低水印的可见度。  2）对水印设计特定的编码器，可以为每张图像产生唯一的ID。  3）将目标检测模型进行对抗训练，用以生成水印系统的编码器。  2、拟采取的研究方法、技术路线、试验方案及其可行性论述  2.1拟采用的研究方法和技术路线   1. 对抗攻击的方法拟采用经典的快速梯度符号方法（FGSM）的攻击方法及其改进的方法。FGSM是一种可以快速实现对抗攻击的单步攻击方法，其优势是速度快，但有比较严重的不足，其攻击的成功率较低且严重影响原始图像的保真度。本课题拟采用基于FGSM提出新的攻击方法解决上述问题。 2. 对抗攻击的目标检测模型拟选用快速检测算法YOLOv3。其特色是能快速的实现目标的检测。   本课题拟采用YOLOv3结合FGSM改进版的对抗攻击方法进行仿真实验，以此展开到其他的目标检测模型的对抗攻击。将通过对以往经典的对抗样本产生方法和进行分析和实验仿真复现，对结果进行比较，筛选出较为合适的方法对目标检测模型分别进行对抗攻击。最后进行评估指标的比较，筛选出最优的结果。  2.2实验方案  1）本课题首先用pytroch复现YOLOv3；  2）提出改进版的FGSM方法：I2-FGSM；  3）采用Microsoft COCO 2014为数据集；  4）采用峰值信噪比（Peak Signal to Noise Ratio，PSNR）和结构相似性（Structural Similarity，SSIM）为评估指标；  5）在YOLOv3利用改进的对抗攻击方法I2-FGSM进行攻击；  6）用PSNR和SSIM评估结果；  7）将I2-FGSM推广到其他模型进行实验，并比较实验结果；  8）设计水印系统的编码器和解码器。  2.3可行性论述  图4 本课题水印系统设计思路  基于对抗样本的产生原理和特性，本课题提出图4的水印系统设计思路。给定一个原始信号S，经过一个嵌入函数G（包括从S中产生对抗样本以及嵌入对抗样本到S中）得到SG，经过设计的编码函数E对产生的水印（即对抗样本）进行编码，得到该水印的唯一ID。如果嵌入此水印的图像被泄露，则可以通过系统设计的独有解码函数D进行解码，得到该图像的对应水印ID及其使用者，达到追溯泄漏源的目标。此设计思路的关键难点在于对抗样本的生成以及编码器和解码器的设计。而本课题已经初步提出了一个对抗攻击的方法，并且在单步目标检测算法YOLOv3中取得了突出的结果。接下来将进行其他目标检测算法的测试和评估。  3、特色与创新性  1）本课题提出的基于对抗样本设计的水印系统是目前已知的对抗样本在水印技术上的首次应用。  2）传统的水印技术存在鲁棒性和感知度无法兼容的冲突，本课题提出的水印系统基于对抗样本，较好的克服了以往水印的不足，并且解决了在面对机器学习算法模型容易被检测和修改的问题。  3）本课题的水印系统设计是目前已知的完全不同于传统水印技术的水印技术，生成和提取水印速度快，并且不需要原图的对比。  4、现有工作基础  1）下载好了实验数据集Microsoft COCO 2014；  2）用pytorch复现了单步目标检测模型YOLOv3；  3）提出了改进版的对抗攻击方法I2-FGSM；  4）将I2-FGSM应用于YOLOv3并取得较为不错的结果，如图5、图6所示。  图5水印系统的工作流程    图6 I2-FGSM在YOLOv3的攻击结果 |

**四、论文工作计划及预期成果**

|  |  |  |
| --- | --- | --- |
| 论文工作进度安排 | | |
| 起止时间 | 工作内容 | 备注 |
| 2019.9.15-  2020.1.20 | 对抗学习领域的接触和了解 |  |
| 2020.2.10-  2020.7.10 | 水印和经典对抗攻击方法的仿真实验对比 |  |
| 2020.7.20-  2021.1.10 | 目标检测算法上的对抗攻击算法实现以及完成开题答辩 |  |
| 2021.1.20-  2021.7.10 | 同时成功对抗攻击两类目标检测算法的实现以及算法的优化，开始论文的撰写 |  |
| 2021.7.20-  2022.1.10 | 水印系统的实现和优化，同时完成论文初稿 |  |
| 2022.1.20-  2022.5 | 论文的完善和答辩 |  |
|  |  |  |
|  |  |  |
| 预期成果  一篇CCF C类以上会议  一项专利  一篇软件著作权  一篇硕士论文 | | |
|  | | |

**五、开题报告作者承诺及导师意见**

|  |
| --- |
| 我的学籍状态为 注册 ，符合学校对开题条件的规定。我保证上述填报内容的真实性，并将在导师指导下，严格遵守学校的有关规定，按计划认真开展硕士学位论文研究工作。  研究生（签名）：  年 月 日 |
| 我已审阅过开题报告的全部内容，同意举行开题报告会。  指导教师（签名）：  年 月 日 |