神经网络水印技术研究进展

作者：张颖君、陈 恺、周赓、吕培卓、刘勇、黄亮 发表时间：2021年

期刊：计算机研究与发展 第五期第58卷

主要内容

- 该文主要阐述分析了水印及其基本需求，并对神经网络水印涉及的相关技术进行介绍；对深度神经网络水印技术进行对比，并重点对白盒和黑盒水印进行详细分析；对神经网络水印攻击技术展开对比，并按照水印攻击目标的不同，对水印鲁棒性攻击、隐蔽性攻击、安全性攻击等技术进行分类介绍；最后对未来方向与挑战进行探讨。
- 该文前面对水印的用途做了大致的讲解：水印的应用主要包括有版权保护、数据监控和数据跟踪等。而随着计算机技术的发展和需要保护的对象的变化和增加，数字水印技术经历了多媒体水印、软件水印到机器学习学习算法\模型水印的发展过程
    - 多媒体水印中常用到的2中关键技术包括扩频水印和量化水印，扩频水印是通过扩频通信技术，将载体信号视为宽带信号，水印信号是为窄带信号，把一个水印的能量谱扩展到很宽的频带中，从而分配到每个频率分量上的水印信号能量较少且难以检测；量化水印是根据水印信息的不同将原始载体数据量化到不同的量化区间，而检测时根据数据所属的量化区间来识别水印信息。
    - 软件水印是解决软件版权问题的重要手段，软件水印主要是指在代码中植入一个特殊的标识符（水印），该水印可以承载软件作者，版权等信息，事后通过特殊的提取器将其从被告软件中识别或抽取出来作为证据以达到检测目的。根据水印的加入位置，软件水印可以分为代码水印和数据水印。
        - 代码水印隐藏在程序的指令部分中，而数据水印则隐藏在包括头文件、字符串和调试信息等数据中．根据水印被加载的方式，软件水印可分为静态水印和动态水印。
        - 静态水印主要是将水印嵌入到可执行程序的代码或数据中，提取过程不需要运行程序，通过静态分析完成识别或提取，主要分为代码替换法、静态图法和抽象解释法等
    - 机器学习方法主要通过统计技术构建数据模型，训练后的模型可以认为是重要的资产，并作为服务MLaaS提供给用户使用，因为需要保护，水印的概念被扩展到机器学习模型领域，嵌入需要保护的模型中以此来保护模型的版权。

神经网络水印技术相关基础

- 当前的机器学习水印主要针对深度神经网络进行研究，包括卷积网络、生成对抗网络以及用于自然语言处理等。
- 该文先是介绍了神经网路后门的相关知识概念，神经网络后门主要是通过在训练集上加上一个或一组特殊的实例（通常被认为是触发器），在神经网络执行分类任务的时候，会执行特殊的分类任务，将特殊实例分类到预设的目标标签中（这通常会违背用户的感知）。在神经网络后门的基础上，我们可以将后门作为水印，实现版权的保护。
- 神经网络后门和基于其的水印都是利用神经网络的过度参数化来学习多个任务，非法用户可以将后门用于恶意目的（例如将驾驶过程中的“停止”标志误分类为“限速”标志），但合法用户可以利用水印防止其部署模型被非法盗用。
- 随后该文简单的对剪枝、知识蒸馏、量化与微调等概念做了解释，并说明了在这些基础上水印也可以进行应用。

神经网络水印技术

基本概念

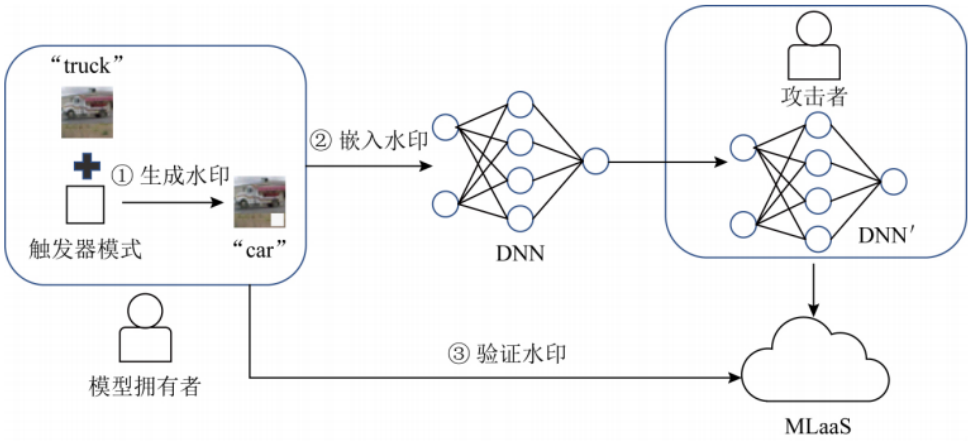
- 神经网络水印添加过程主要是通过在模型中添加一个额外的训练目标来注入水印，并且要尽可能满足保真度、安全性、鲁棒性和隐蔽性等需求。过程包括有生成水印、嵌入水印和验证水印。
    - 生成水印是指模型所有者设计特殊的水印形式，例如一个比特串或者一些经过特殊设计的训练样本，以便模型在验证水印阶段能够以某种特殊的方式验证水印的存在性。
    - 嵌入水印是指将生成的水印信息插入到神经网络模型中。
    - 在验证水印时，需要输入特定数据，然后观察模型的反馈或者输出，与预期结果进行匹配，从而验证水印的存在性。
    - 

Fig. 5　Three steps to DNN watermark

图 5　DNN 水印 3 个步骤

- 深度神经网络水印技术的对比
    - 表 2　DNN 水印技术对比

参考文献	年份	水印方法	分类	水印大小	嵌入方式	方法说明
文献[21]	2017	将比特数组作为水印嵌入神经网络某层的权重	白盒	多比特	训练	适用于多种神经网络，能抵御剪枝、微调的攻击
文献[31]	2018	将抽象图片作为后门水印嵌入神经网络，后门结合了加密方法保证了所有权	黑盒	0 比特	训练/微调	能应对剪枝、微调、蒸馏攻击
文献[39]	2019	根据所有者的签名生成水印密钥对	黑盒	多比特	微调	能够抵御剪枝、微调、水印覆盖攻击
文献[40]	2019	将所有者的水印嵌入在模型的不同层中获得的数据抽象(特征输出)的概率密度函数中	白盒/黑盒	多比特	训练	能够抵御多种移除和转换攻击，包括模型压缩、模型微调和水印覆盖
文献[41]	2020	用类似 GAN 的方法，使得带水印的参数相似于不带水印的网络	白盒	多比特	训练	可以抵抗属性推断攻击和水印移除
文献[42]	2018	将指纹嵌入到训练好的模型的权重概率密度函数中	白盒	多比特	训练	能够应对指纹勾结、剪枝、微调攻击
文献[43]	2018	提供了基于文本、噪声、不相关图片作为触发器的后门水印方式	黑盒	0 比特	训练/微调	能应对剪枝、微调、蒸馏攻击
文献[44]	2018	提出了一种将由比特数组生成的特定信息作为触发器的后门水印	黑盒	多比特	微调	能够抵御模型篡改；水印具有隐蔽性，攻击者很难发现水印，并恶意声称所有权
文献[45]	2020	在训练过程中为关键样本增加新标签	黑盒	0 比特	微调	假阳性率低，能抵抗剪枝、微调攻击
文献[46]	2019	提出了一种指数加权的后门水印方法	黑盒	0 比特	训练	将后门水印对参数的影响施加在较大值的权重参数上(指数加权实现)，保证水印在剪枝微调更具鲁棒性
文献[47]	2020	部署在模型的预测 API 中，通过改变客户端预测响应来动态地为部分查询添加水印	黑盒	0 比特	微调	能有效抵抗模型提取以及分布式提取攻击
文献[48]	2020	训练前根据图案将原始图像像素更改为正值或负值	黑盒	0 比特	微调	能抵抗剪枝微调、模型提取攻击，有效防止水印检测和移除，并且对迁移学习有很好的适应性
文献[49]	2019	提出了一种通过编码器生成盲后门水印方法	黑盒	0 比特	训练	由于这是一种盲水印，能够轻易的躲避人肉眼的检测，更具隐蔽性；能够抵御剪枝、微调的攻击
文献[50]	2020	利用对抗样本刻画决策边界，并以此作为水印	黑盒	0 比特	微调	能够抵御剪枝、微调、奇异值分解的攻击
文献[51]	2019	基于进化算法生成和优化触发集	黑盒	0 比特	训练	假阳性率低，能抵抗微调攻击
文献[52]	2019	在网络结构上插入一层数字护照层	白盒/黑盒	多比特	训练	使用配套护照才可以正常使用神经网络；护照被盗可以通过签名验证所有权

- 其中根据DNN模型是否公开，白盒水印和黑盒水印2类进行介绍和对比分析。白盒水印和黑盒水印主要是根据在水印插入和验证时是否需要获取模型本身进行划分．具体地，白盒水印是指需要获取模型相关参数；黑盒水印是指水印的执行过程不需要访问模型本身，该方法主要通过机器学习服务中API对黑盒水印进行提取测试。
    - 白盒水印是将生成的水印嵌入到DNN模型参数中，然后从模型提取标记进行验证。
    - 黑盒水印由于白盒水印在验证环节需要模型拥有者知道可疑模型的内部细节（如结构、参数等），才能提取其完整水印，并与嵌入的水印对比位错误来完成验证，因此适用性受到了很大限制．因此，有学者提出了以黑盒的方式为模型添加水印的方法，从而在无需知晓模型参数等细节的情况下进行水印的验证。

神经网络水印攻击方法

对鲁棒性的攻击

- 鲁棒性指水印不易去除。如强行去除后会损坏模型的保真度。攻击者的目标是在保持一定保真度的情况下，使得模型水印失去其作用，即模型拥有者O无法确认模型中水印的存在。一个简单的方法去除模型中的水印即重训练一个新的模型，但该方法需要大量的训练数据和计算能力．若攻击者有此能力，无需再剽窃其他人的水印．因此攻击者尝试使用少量的训练样本甚至不使用训练样本进行水印的去除。

对隐蔽性的攻击

- 隐蔽性指水印隐藏在模型中，不易被发现，通过隐蔽性达到不易去除的效果．从攻击者角度，可尝试发现水印，从而移除水印或者宣称对水印的所有权。隐蔽性水印的设计被用来防止此类攻击．

对安全性的攻击

- 安全性指水印本身不易伪造，否则攻击者也能伪造水印从而宣称对嵌入水印模型的所有权．这里有2类方法：1)攻击者尝试发现 W 的构造，从而宣称拥有 W。2)攻击者再次嵌入类似的水印W'到Mw中，得到含 2 个水印的模型 Mw + W'，从而宣称对 M 的所有权．

其他方法

总结

- 该文指出，从目前来看，机器学习模型水印技术还处于发展前期，在理论上和实际应用上并不完善，攻击者仍然有多种方法能够对已有的保护方法进行攻击。并给出了未来在5各方面值得探索
    - 更为鲁棒的神经网络模型水印
    - 水印应减少对原始模型的影响
    - 公开的水印
    - 水印的理论证明
    - 多样的水印
- 其次，自己阅读之后也有一些收获，在未阅读之前，以为本文讲的是一种使用神经网络方法的水印，但不曾想到是对神经网络模型添加水印，这为我打开一个新的方向，水印是保护产品版权问题所存在的，现在计算机技术的发展，带来了不少的新兴技术，也引发了更多的安全问题，信息安全对于现在的生活是急需品，每个人都生活在信息时代，应当好好考虑如何才能保证信息安全和技术安全。水印也应该要扩宽其业务，与更多的领域进行交叉融合。