人工智能模型水印研究综述

作者：谢宸琪 张保稳 易 平 时间：2021年

期刊：计算机科学 第7期第48卷

主要内容：

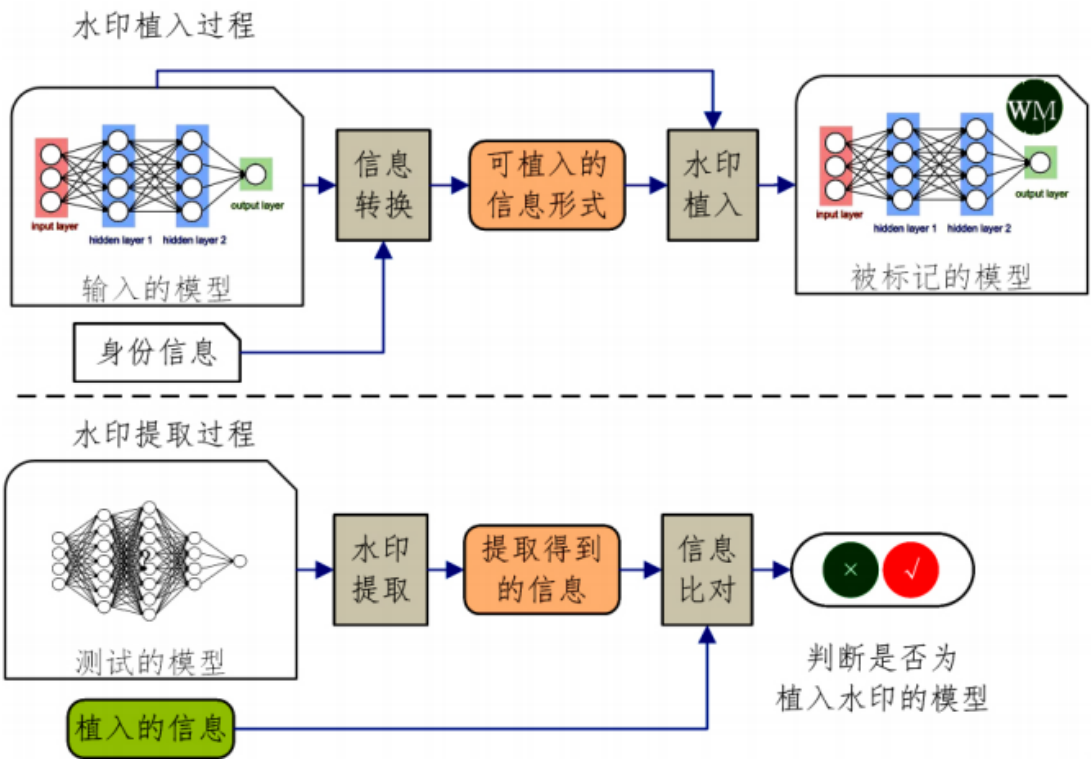
- 该文与上次阅读的神经网络水印的发展研究进程一文所表达的内容大致一致，都是针对现目前人工智能模型进行水印的添加，从而达到对模型的保护。
- 该文主要将对目前的模型水印研究现状做一个全面的介绍，按照其生成方式和原理的不同进行分类，介绍其原理并分析各种模型水印算法的优缺点，在此基础上探讨模型水印的研究方向和未来的发展趋势。

1.模型水印

1.1 概念

- Uchida等首次尝试向深度神经网络添加水印。模型水印利用数字水印技术，通过某种通用框架向模型中植入水印，以检测训练好的模型是否被侵权的技术

1.2基本流程操作

- 

2. 模型水印相关分析

2.1模型水印的存在性原理

- 我们知道将水印嵌入到图像中其实是利用图像对于人眼的像素冗余空间来实现的。因为人眼对每个元素的细微变化并不敏感，也就是说对于人眼来说图片像素的细微变化不会引起视觉上的较大差异，这一部分像素空间的可变化区间便是对于人眼的看图冗余。
- 而模型水印其实也是根据这一原理来实现的。模型会承载大量的数据信息，对于模型的能力来说，同一个分类问题，模型参数的细微变化，对模型的分类准确率并不会造成明显影响，因此这一部分的参数空间的可变化区间就是模型参数空间的冗余。因此我们可以将信息嵌入到这一区间中，从而实现模型水印的预期功能

3.模型水印研究现状

3.1 修改训练数据加水印的方法

- 利用后门植入水印的方法
- 利用对抗样本构建水印的方法

3.2修改模型加水印的方法

- 利用投影矩阵构建水印的方法
  - 水印嵌入

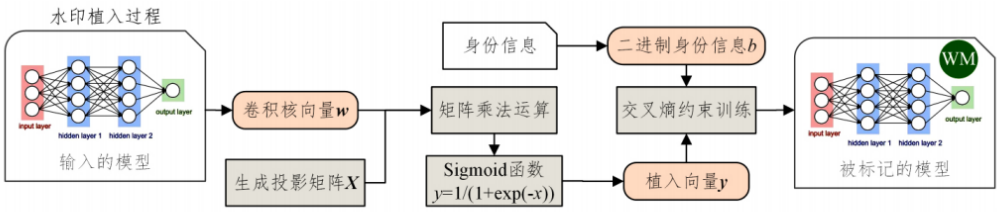
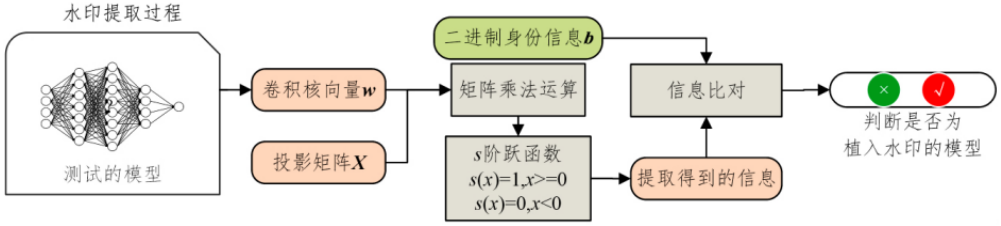


图 3　水印植入流程图

  - 水印提取



- 添加模型结构构建水印的方法
- 利用聚类构建水印的方法
- 利用对抗网络训练的方法

4.水印算法比较

水印算法	优点	缺点
后门植入水印	验证方便，只需要输入特定图片，就可以进行黑盒提取	无法抵御预测 APIs 的模型窃取攻击
利用对抗样本构建水印	植入水印后可提高对于对抗样本的鲁棒性，可进行黑盒提取	对抗样本在不同模型上的迁移性不确定
利用投影矩阵构建水印	过程相对简单，无须添加网络结构	无法抵抗水印覆写
在模型中添加结构	对水印覆写具有一定的鲁棒性	模型结构相对复杂，参数量增加
利用聚类将图片按输出激活分类编码	能够提升模型对于对抗样本的鲁棒性	密钥集构建烦琐，过程相对复杂
利用对抗网络训练	输入图片的分布与正常图片相近，不易被筛出	多个网络同时进行训练，工作量大

5.结论

- 该文最后指出，模型水印算法目前主要集中在对模型本省植入水印的方式，而模型参数中包含着大量的冗余，有足够的冗余空间给我们植入水印，并且我们可以结合信息论、编码理论、密码学等，以寻找不同的方式来实现对水印的植入。
- 对于模型水印的性能，目前评估的指标并不明确，只有少数几个指标是公认的，但由于不同水印算法采用的训练方式不同，考察的指标不同、实验参数不同等，导致不同研究成果之间难以做出比较，因此对模型水印性能评估的分析需要一个统一的框架和标准

6. 总结

- 阅读该文，系统性的了解到了针对模型这一非常有价值的知识产权资产进行添加水印的方法以及他们的相关概念，补充了之前阅读上一篇论文的其他知识面