深度学习模型版权保护平台

组长：张卓萌

小组成员：章杭炜 李佳露 罗书卿 蔡锶维 朱文骏 张昊

需求分析

背景概述

近年来，伴随着计算机存储设备、处理器算力、问题求解算法等各个方面的不断发展，以深度学习为代表的人工智能技术出现重要突破并在社会生活中的很多方面得到广泛应用，有力驱动了产业升级和科技进步。但是，训练出一个实用的深度学习模型并不是一件容易的事情，这需要在**数据集、神经网络结构、处理器算力**等诸多方面做出很大的投入。

用户或者企业投入大量的人力物力财力训练出可以实际应用的深度学习模型之后，如果该模型被投入商业化应用，那么**我们有必要采取某种方法来证明该用户对这一模型的所有权，以防止有模型盗窃者窃取别人的模型用来给己方牟利。**

随着深度学习技术的广泛应用，深度学习模型的版权保护技术也逐渐受到世界各国的关注。例如，2017 年 7 月，中国国务院印发了《新一代人工智能发展规划》，强调要建立人工智能技术标准和知识产权体系。2018 年 11 月，欧洲专利局发布了人工智能和机器学习的专利性指南。由此可见，**如何有效解决深度学习模型的版权保护问题，规范模型的所有权认证方式，已经是箭在弦上了。**

基于当下模型商业化的趋势和诸多现实的需求，我们旨在搭建一个**基于水印算法的深度学习模型版权保护平台，为需要对模型实施版权保护的用户提供高效的水印加注方案，并且能够对侵犯版权的行为做出客观的裁决。**

算法功能需求

作为一种特殊的数据形式，深度学习模型的版权保护和普通数据的版权保护要求相比有一定的不同之处，具体可总结为以下四个方面：

- **功能不变性**：这也是模型版权保护最基本的要求，即嵌入了水印的模型相对于原始的模型，在功能上不能有太大的变动。这里的功能体现为识别的准确率等评估模型性能的指标。也就是说，版权保护方法不能以牺牲模型功能为代价，否则就失去了它的意义。
- **鲁棒性**：模型保护方法需要能抵抗多种攻击手段，比如模型微调、剪枝、神经元置乱、压缩等。
- **安全性**：和密码学中的柯克霍夫原则类似，深度学习模型的版权保护也应该遵循同样的准则，即每个模型采取的版权保护算法都是公开的，对于资源有限的侵权者，只要他没能获取到密钥，就无法对模型实施有效的侵权或者攻击。
- **计算复杂度**：深度学习模型版权保护算法的计算复杂度可以被评估，而且不能过高。

平台功能需求

- **基本用户**

  > - 用户注册
  > - 用户登录
  > - 个人信息的查询和修改

- **管理员**

    - 平台情况的监测
    - 推荐算法的修改和拓展

- **模型注册**

    - 模型注册申请
    - 多场景应用，算法推荐
    - 详情查询
    - 数据下载
    - 材料提交

- **模型裁决**

    - 多场景应用
    - 黑盒 / 白盒裁决