**Privacy Risks of Securing Machine Learning Models against Adversarial Examples**

**背景：**近年来，对机器学习模型的攻击和防御逐渐火热，尤其是在安全和隐私领域。但是，先前研究的很大限制是安全领域和隐私领域都是相对独立分开的。一个领域的防御方法是否会对另外一个领域产生影响目前还不得而知。因此，这篇文章结合两方面综合考虑防御的性能。在安全领域，敌手的目标是诱导对目标机器学习模型进行错误分类。攻击方法为逃逸攻击和毒药攻击。前者是在测试时干扰输入，从而导致目标模型错误预测。后者则是恶意修改部分训练数据来控制训练过程，导致训练模型在某些测试输入上错误表现。为了应对这两种攻击，安全领域设计了新的训练算法，以确保机器学习模型避免逃逸攻击和毒药攻击。而在隐私领域，敌手的目标是获取关于模型的训练数据或目标模型的私人信息。攻击类型有：针对数据隐私的成员关系推理攻击、学习训练数据全局属性的属性推理攻击、隐信道模型训练攻击。攻击手段有：模型提取攻击、超参数偷取攻击。

**摘要：**衡量了六种最先进的防御方法对成员关系推理攻击的成功程度

**贡献：**

1. 利用对抗样本的预测结果和最坏验证的预测结果提出了两种新的成员关系推理攻击，可以得到更高置信度的推理精度。

2. 使用六种最先进的对抗防御方法进行成员关系推理攻击。证明所有方法都会增加模型成员关系推理风险。

3. 针对敌手防御性强模型，讨论了影响成员关系推理性能的因素，包括鲁棒泛化性、对抗扰动性的约束以及模型容量

4. 实验评估敌手先验知识、防御措施和正则化的影响，并讨论了训练数据隐私和模型鲁棒性之间的关系

物理在良性还是对抗的环境下，由于表现出更大的泛化误差，鲁棒模型可能会泄露更多的成员信息。而且，鲁棒训练算法可能通过增加模型对训练数据的敏感性，使模型更容易受到成员推理攻击。

