《人工智能安全》课程作业

HW 3：人工智能隐私性 课后思考题

姓名： 周炜

学号： 3210103790

专业： 计算机科学与技术

邮箱： 3210103790@zju.edu.cn

Question 1

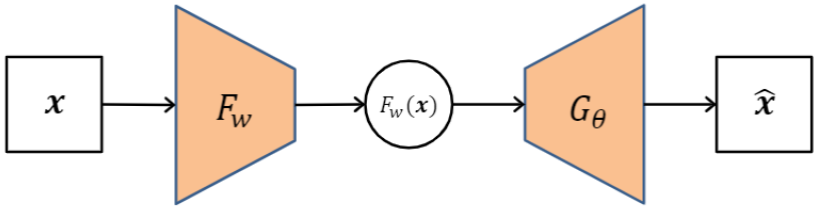
> 白盒模型逆向攻击和黑盒模型反演攻击各适用于什么样的场景？课程中的模型逆向攻击都需要最终的输出向量，是否有可能进行仅获得标签的模型逆向攻击？

白盒模型逆向攻击

攻击者知道模型结构、 参数。有了这些特权信息，攻击者可以执行逆向过程，重建原始训练数据或模型参数的近似值。当攻击者拥有关于模型的全面知识时，白盒攻击是最有效的，例如可以访问其源代码或有能力检查其内部状态

黑盒模型反演攻击

攻击者只能访问目标模型, 得到输出向量, 并不知道模型的结构、参数 。攻击者作为一个外部实体与模型互动，提交输入查询并观察相应的输出



黑盒模型逆向攻击原理示意图

进行仅获得标签的模型逆向攻击**非常困难**（或者说是不太行）。因为在典型的模型逆向攻击中，模型的最终输出向量或概率的可用性对于指导逆向过程和准确估计原始训练数据或模型参数至关重要。如果不能获得最终的输出向量或概率，进行仅有标签的模型反转攻击就变得非常困难。没有明确的指导，阻碍了训练数据的有意义的表征的重建，使得实现成功的反演变得很没有意义

Question 2

> 模型窃取攻击中，替代模型方法异常的大量查询不仅仅会增加窃取成本，更会被模型拥有者检测出来，你能想到什么解决方法来避免过多的向目标模型查询？

抽样查询

对提供最有价值的信息的查询子集进行采样来替代全部采样。通过仔细选择涵盖模型决策边界和关键特征的代表性输入，同时最大限度地减少整体查询量和相关的检测风险

查询优化

冗余或高度相关的查询可以被识别并从攻击过程中消除。此外，攻击者可以优先考虑可能产生更多信息的查询，如位于决策边界的输入或与模型预测的高不确定性相关的输入

转移学习

在一个单独的数据集上训练一个替代模型，以捕捉目标模型的行为。通过查询代用模型或者是辅助模型，攻击者可以减少对目标模型的过度查询，同时仍然获得有价值的信息

自适应查询

可以采用自适应主动学习的方法，智能地选择信息量最大的查询，从而减少整个查询次数。通过根据模型的不确定性反复选择查询，或将重点放在高度重要的领域，攻击者可以自适应地查询目标模型，有效减少所需的查询次数，同时最大限度地获得知识

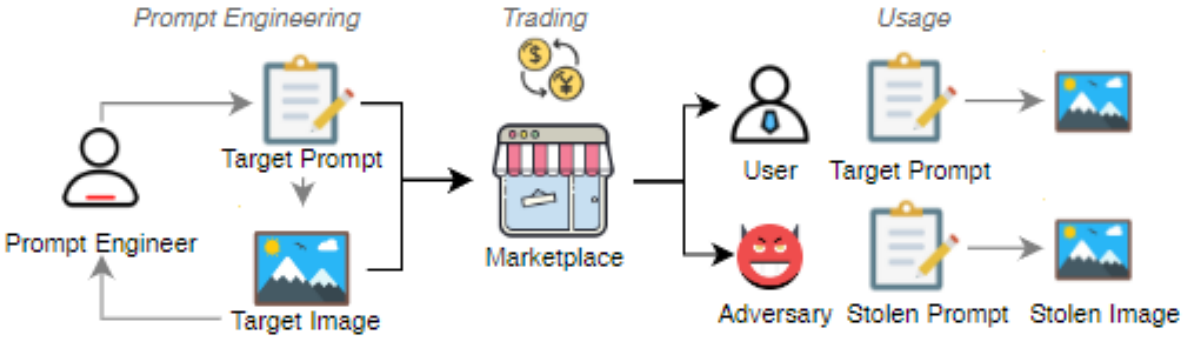
利用PCA等聚类算法

利用聚类算法筛选具有代表性的"中心"样本进行查询

针对特定攻击目标的优化

比如如果要窃取**AIGC**模型，则需要在如何涉及**prompt**上下功夫(一种新的工作类型，即提示工程师，已经出现了,专注于制作高质量的提示。 还有，高品质提示成为新的有价值的商品，并且在专业市场交易，例如 PromptBase ， PromptSea 和 Visualize AI)，下列以最近**ICML**大火的`stable diffusion model`为例

参考论文 [Prompt Stealing Attacks Against Text-to-Image Generation Models]，可以用**提示窃取攻击**，通过文本到图像生成模型从生成的图像中窃取生成图片和它对应的高效提示，然后可以根据这组窃取的数据，进行相似模型的训练，从而达到窃取模型的效果



减少被发现风险的方法

差分隐私

可以采用差分隐私来保护攻击者所做查询的隐私。通过在查询中添加精心校准的噪音，攻击者可以确保查询不会暴露关于目标模型或用于训练模型的数据的敏感信息。这可以使防御者更难区分合法的用户查询和攻击者发起的查询

模仿正常查询

攻击者可以通过模仿合法用户行为中常见的正常查询模式来减少怀疑，混杂攻击查询和正常查询，以降低被发现的风险

查询随机化

在查询过程中引入随机化技术（或者说是引入噪声）——通过改变查询的时间、数量和内容，攻击者可以使他们的活动不那么可预测，类似于用户行为的自然变化。但是，主要注意，必需在随机化和保持被盗模型的保真度之间取得平衡

匿名化机制

利用匿名化技术来混淆他们的身份并隐藏他们的真实意图。比如，用匿名代理或虚拟私人网络（VPN）来掩盖他们的IP地址，使防御者难以追踪到查询的来源，从而找到攻击者。或者经常切换IP以及用户来进行攻击

Question 3

> 在MemGuard 的防御场景下，如果攻击者在输入图像上添加扰动可以破坏单次随机的设定，你认为防御者应该如何应对?

> 参考论文：MemGuard: Defending against Black-Box Membership Inference Attacks via Adversarial Examples

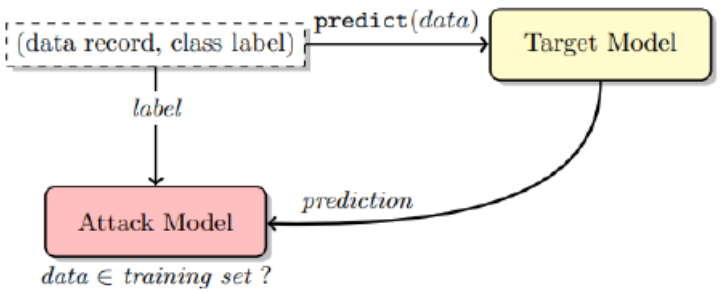
图像预处理

使用滤波等手段进行图像预处理，以检测和减少去除添加的扰动。这些技术可能涉及复杂的算法，利用机器学习模型来区分合法的图像内容和对抗性扰动。通过识别和消除扰动，防御者可以确保保留一次性随机设置并保持系统的完整性

对抗性训练

这种技术涉及到用对抗性例子来增强训练过程，以提高模型对扰动的鲁棒性。通过在训练过程中让模型暴露在扰动的图像中，它在推理过程中更善于识别和忽略对抗性扰动，从而保留了一次性随机设置

上课提到的一种方法就是可以采用对抗样本的思想，通过向在置信度向量中 添加微小的噪音 扰动 使原置信度向量移动到攻击者的分类器的决策边界 从而使攻击者无法从置信度向量分类出该样本是否属于原训练数据



模糊哈希算法

使用模糊哈希算法，根据特定条件对文件进行分段，然后使用鲁棒哈希算法计算每个分段的哈希值。 这些值的一个子集被选择并连接起来以形成包含分割条件的模糊散列结果。 这样不仅可以压缩输入内容，而且保留了两个文件之间的相似度。因此，较小的扰动将对模糊散列结果产生最小的影响

异常检测和手动查验

利用异常检测算法或雇用人类专家手动检查可疑的图像或输入数据

Question 4

> 数字水印可以保护模型版权，但是无法防御攻击者窃取模型的过程，是否有方法可以直接防止模型被窃取?

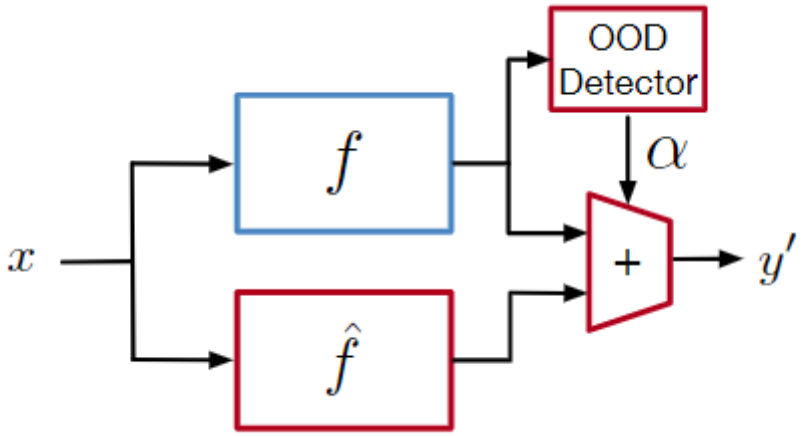
这方面网上的资料较少，因此我选择看了一些论文，以下四种方法是比较新或者比较有影响力的方法

自适应错误信息干扰 Adaptive Misinformation

> 参考论文：Defending Against Model Stealing Attacks With Adaptive Misinformation

作者发现现有的模型窃取攻击总是使用**OOD的数据**来查询目标模型。因此可以通过识别OOD的输入并使用错误信息选择性地为此类输入提供服务来干扰窃取模型攻击

如下图，作者使用 OOD 检测机制选择性地使用错误信息函数f'的预测来服务 OOD 输入，而 ID 输入使用正确模型f的原始预测来服务



欺骗性扰动 Deceptive Perturbations

> 参考论文：Defending Against Neural Network Model Stealing Attacks Using Deceptive Perturbations

作者提出了一个可以应用于大多数神经网络分类器的附加层，该层增加了一个小的可控扰动，在保持准确性的同时，最大限度地减少被盗模型的损失。也就是说，并不是试图检测攻击，而是应用对正常用户影响不大的噪音，但仍然会降低和减缓模型窃取攻击。一个最佳的防御应该为消费者提供服务，同时除了最后的标签之外，不给对手提供任何可衡量的好处

从下图可以看到，运用不哦那个的函数和策略在附加层中，防御效果不一样，需要对特定的网络设置特定的附加层

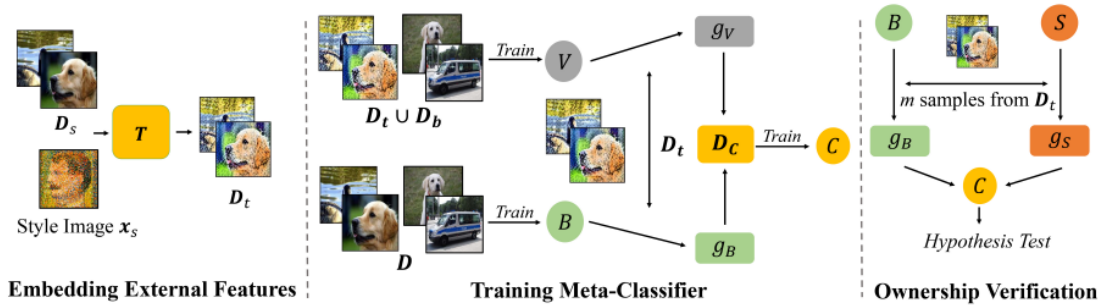
Figure 5: Agreement of stolen model with base model using different defense types.

风格迁移 Verifying Embedded External Features

> 参考论文：[Defending against Model Stealing via Verifying Embedded External Features]

在本文中，作者将模型窃取的防御表述为验证可疑模型是否包含防御者指定的外部特征的知识。 具体来说，可以通过使用**风格迁移**修改一些训练样本来嵌入外部特征。 这种方法的灵感来自于这样一种理解，即被盗模型应该包含受害者模型学习到的特征知识。并且，作者在 CIFAR-10 和 ImageNet 数据集上评估了我们的防御，验证了我们的方法可以同时防御各种类型的模型窃取，同时保持预测良性样本的高精度

作者提供的方法主要流程为：在第一阶段，防御者将通过风格转移修改一些图像，以嵌入外部特征。在第二阶段，防御者将训练一个元分类器，以确定一个可疑的模型是否是基于梯度从受害者那里偷来的。在最后阶段，防御者将通过假设测试进行所有权验证。如下图所示



作者还将他们的防御与数据集推理和模型水印与 BadNets、梯度匹配 和纠缠水印三种方法进行了比较，不同防御涉及的图像示例如下图所示

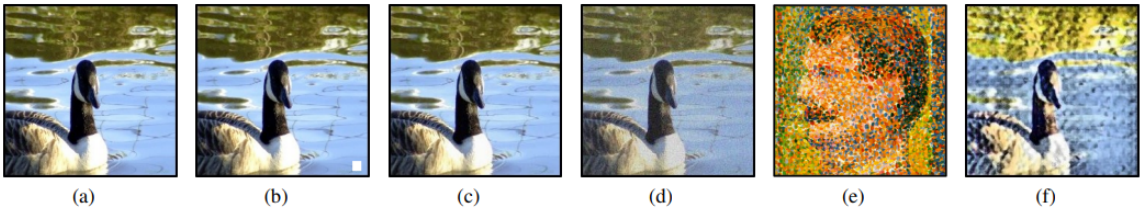


Figure 4: Images involved in different defenses. **(a)** benign image; **(b)** poisoned image in BadNets; **(c)** poisoned image in Gradient Matching; **(d)** poisoned image in Entangled Watermarks; **(e)** style image; **(f)** transformed image.

Model Stealing		BadNets		Gradient Matching		Entangled Watermarks		Dataset Inference		Ours	
		$\Delta\mu$	p-value	$\Delta\mu$	p-value	$\Delta\mu$	p-value	$\Delta\mu$	p-value	$\Delta\mu$	p-value
Victim	Source	0.91	10^{-12}	0.88	10^{-12}	**0.99**	$\mathbf{10^{-35}}$	-	10^{-4}	0.97	10^{-7}
$\mathcal{A}_D$	Distillation	-10^{-3}	0.32	10^{-7}	0.20	0.01	0.33	-	10^{-4}	**0.53**	$\mathbf{10^{-7}}$
$\mathcal{A}_M$	Zero-shot	10^{-25}	0.22	10^{-24}	0.22	10^{-3}	10^{-3}	-	10^{-2}	**0.52**	$\mathbf{10^{-5}}$
	Fine-tuning	10^{-23}	0.28	10^{-27}	0.28	0.35	0.01	-	10^{-5}	**0.50**	$\mathbf{10^{-6}}$
$\mathcal{A}_Q$	Label-query	10^{-27}	0.20	10^{-30}	0.34	10^{-5}	0.62	-	10^{-3}	**0.52**	$\mathbf{10^{-4}}$
	Logit-query	10^{-27}	0.23	10^{-23}	0.33	10^{-6}	0.64	-	10^{-3}	**0.54**	$\mathbf{10^{-4}}$
Benign	Independent	10^{-20}	0.33	10^{-12}	0.99	10^{-22}	0.68	-	**1.00**	**0.00**	**1.00**

Table 3: Results on CIFAR-10 dataset.

Model Stealing		BadNets		Gradient Matching		Entangled Watermarks		Dataset Inference		Ours	
		$\Delta\mu$	p-value	$\Delta\mu$	p-value	$\Delta\mu$	p-value	$\Delta\mu$	p-value	$\Delta\mu$	p-value
Victim	Source	0.87	10^{-10}	0.77	10^{-10}	**0.99**	$\mathbf{10^{-25}}$	-	10^{-6}	0.90	10^{-5}
$\mathcal{A}_D$	Distillation	10^{-4}	0.43	10^{-12}	0.43	10^{-6}	0.19	-	10^{-3}	**0.61**	$\mathbf{10^{-5}}$
$\mathcal{A}_M$	Zero-shot	10^{-12}	0.33	10^{-18}	0.43	10^{-3}	0.46	-	10^{-3}	**0.53**	$\mathbf{10^{-4}}$
	Fine-tuning	10^{-20}	0.20	10^{-12}	0.47	0.46	0.01	-	10^{-4}	**0.60**	$\mathbf{10^{-5}}$
$\mathcal{A}_Q$	Label-query	10^{-23}	0.29	10^{-22}	0.50	10^{-7}	0.45	-	$\mathbf{10^{-3}}$	**0.55**	$\mathbf{10^{-3}}$
	Logit-query	10^{-23}	0.38	10^{-12}	0.22	10^{-6}	0.36	-	10^{-3}	**0.55**	$\mathbf{10^{-4}}$
Benign	Independent	10^{-24}	0.38	10^{-23}	0.78	$\mathbf{10^{-30}}$	0.55	-	0.98	10^{-5}	**0.99**

Table 4: Results on ImageNet dataset.

如上表所示，作者提供的防御方法在几乎所有情况下都达到了最佳性能

梯度限制 gradient redirection

> 参考论文：Model Stealing Defenses with Gradient Redirection

作者提出了一种可证明的最优算法来有效地解决梯度重定向问题，并且用它来构建 $GRAD^2$ 模型窃取防御(较为复杂，不再赘述，可以参见论文)

如下图，用户将留下的图像提交给预测 API，生成后验概率 y。 我们的防御输出经过调整的后验眼，以防止恶意用户窃取模型（红色，右）。 调整是由一种有效的、可证明是最优的算法选择的，以在代理网络的梯度中产生最大的误差，该梯度转移到未知的对手网络。 大量实验表明，被盗模型在准确性上有显着损失。 此外，善意的用户影响不会很大，因为保证调整很小