后门攻击思考题 课后思考

问题一

对抗样本、数据投毒、后门攻击本质上的区别是什么？它们分别适合现实中什么样的场景？

我认为，对抗样本、数据投毒、后门攻击区别在于它们的**攻击方式和攻击目的不同。**

对抗样本：是指通过对输入数据进行微小的修改来欺骗人工智能模型，使其产生错误的输出。主要目的是破坏模型的准确性，适用于需要误导人工智能模型的场景，比如图像分类、语音识别等，家作中就做过设计mnist的对抗数据，以此来让模型的准确度下降。

数据投毒：是指在训练数据中注入恶意数据，以改变人工智能模型的训练结果。主要目的是破坏模型的**训练过程的数据安全**，适用于需要破坏人工智能模型训练的场景，比如垃圾邮件过滤、恶意软件检测等。课堂上我们曾举了设计大量恶意数据区攻击购物系统，使得模型的训练效果出错。

后门攻击：是向模型中植入恶意代码，在特定条件下可以控制或破坏模型，主要适用于控制或破坏模型的场景，破坏**训练过程中的模型安全。**主要目的是控制或破坏人工智能模型，适用于需要控制或破坏人工智能模型的场景，比如机器学习系统的身份认证、安全检测等。课堂上我们讲过，开发者通过设定特殊的规则，使得一触发规则就造成结果的巨大差异。

而在于**应用场景**方面

对抗样本：主要应用于图像识别、语音识别、自然语言处理等领域。这些领域的人工智能系统普遍需要输入数据并进行分类或者预测，对抗样本可以通过对数据进行相应的修改来欺骗模型，破坏模型的准确性。

数据投毒：主要用于攻击具有分类任务的人工智能模型，如垃圾邮件过滤、恶意软件检测等。数据投毒的攻击方式中，攻击者批量制造恶意数据接注入到训练集中，在不被发现的情况下改变模型的训练结果。

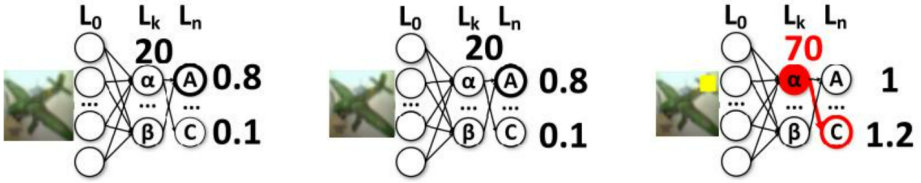
后门攻击：主要用于攻击需要验证的人工智能模型，如安全检测等。后门攻击可以在模型中植入恶意代码，使得攻击者可以在特定条件下控制或破坏模型，一旦触发器条件被激活，就会大大威胁模型的安全性。

问题二

白盒模型后门检测和黑盒模型后门检测各有什么优缺点？它们分别适合现实中怎样的场景

白盒模型后门检测：

- 优点：
  - 可深入：能够深入分析模型的内部结构和算法，从而可以检测到一些深度隐藏的后门
  - 准确性高：白盒模型后门检测可以查看模型内部的结构和参数，对于被植入后门的模型特征进行全面的分析和检测，下面的图里就列举了课上讲过的检测受损的神经元的思想
  - 适用范围广：白盒模型后门检测不需要对攻击方式进行假设或猜测，能够较好地适应各种类型的后门攻击



(a) Benign model with benign image (b) Trojaned model with benign image (c) Trojaned model with trigger image

- 缺点：
  - 实现困难：需要获取模型的源代码和参数信息，这在实践中难以实现。有些模型检测如果不是内部检测，可能无法给检测方
  - 耗费资源较多：为了进行深度的白盒后门检测，需要监控模型的所有输入和输出以及大模型，这可能会需要大量运行时间和存储空间
  - 对模型产生修改：为了弥补问题，可能会模型参数进行修改，这之中会影响到模型的准确度

黑盒模型后门检测:

- 优点
  - 易于开展：不需要获取模型的源代码和参数信息，只需要利用输入和输出进行检测，因此比较容易实现。
  - 不会影响模型：不会对模型进行修改，可以确保不影响到大部分样本的准确度，可以保证模型不会因为检测而产生较大变化
- 缺点
  - 检测难以深入：无法分析模型的内部结构和算法，无法检测到一些深层次的后门。
  - 准确性较低：黑盒后门检测只能根据模型的输入和输出进行分析，很难发现一些深度隐藏的后门、或者需要多步操作才能触发的后门。但这并不绝对，比如课堂上讲过的基于元学习的方法，就甚至超过了白盒检测

Approach	MNIST-M	MNIST-B	CIFAR10-M	CIFAR10-B	SC-M	SC-B	Irish-M	Irish-B	MR-M
AC [12]	73.27%	78.61%	85.99%	74.62%	79.69%	82.86%	56.14%	93.48%	88.26%
NC [53]	92.43%	89.94%	53.71%	57.23%	91.21%	96.68%	✗	✗	✗
Spectral [52]	56.08%	≤50%	88.37%	58.64%	≤50%	≤50%	56.50%	≤50%	**95.70%**
STRIP [21]	85.06%	66.11%	85.55%	81.45%	89.84%	85.94%	≤50%	≤50%	✗
MNTD (Jumbo)	**99.77%**	**99.99%**	**91.95%**	**95.45%**	**99.90%**	**99.83%**	**98.10%**	**99.98%**	89.23%

白盒模型后门检测适用于：

- 模型的安全需求较高的场景；
- 可以获得模型源代码和参数信息的情况下，最好是内部人员自己检测；
- 需要进行深度检查，确保所有可能的后门都被发现的场景；
- 需要对已知攻击类型进行检测的场景。

黑盒模型后门检测适用于：

- 无法获得模型源代码和参数信息的情况下；
- 需要对多个模型进行**快速**检测的场景；
- 可以使用各种类型的数据集进行测试的场景；
- 需要对未知攻击类型进行检测的场景。

问题三

为了防范模型后门攻击，你还能想到什么样的方法来提前预防这类恶意攻击对模型的可用性和完整性产生破坏?

- 安全开发规范：**训练之前**，使用安全编码标准、漏洞扫描工具和安全测试工具，建立安全开发流程和制定相应的安全政策和规范，确保整个开发过程的安全性和可追溯性，以减小有人破坏模型的可能性动机。

- 训练数据检测：**训练之前**，还可以使用数据清洗和预处理技术来减少后门注入的可能性。例如，可以使用去噪和去重技术来减少冗余和无效数据，可以使用数据聚类和分类算法来判断数据是否符合正常分布，并及时删除异常数据。
- 训练数据增强：**训练之中，**可以设计一些特定的攻击数据混入其中，来提高模型的鲁棒性，不至于很容易被攻击
- 限制模型访问权限：**训练之中**，使用身份认证和授权技术，并在模型部署时设置多层次的访问权限，例如，根据用户的角色和职责设置不同的访问级别和权限。此外，还可以为模型建立操作审计日志，并定期对日志进行分析和审计。以断绝训练时有人恶意注入后门
- 设置安全密钥：**训练之中**，为了确保模型在训练和部署过程中不会被篡改或污染，可以采取一些安全措施来加强模型的安全性，如加密、签名、验证等。例如，可以使用加密技术对模型进行加密，只有有权访问密钥的人才可以进行解密。此外，可以使用数字签名对模型进行签名，以确保没有人对模型进行了更改。
- 模型监控：**训练之中**，使用异常检测技术和事务管理器，使用流量分析和行为分析技术来监控模型的输入和输出，及时发现和处理异常情况。此外，还可以使用机器学习技术来自动化模型的监控和管理，从而减轻人工干预的压力。
- 多模型检测：**训练之中**，还可以使用模型退化技术（model degradation）来掩盖模型的关键信息，从而降低模型被攻击的风险。

- 模型评估：**训练之中**，我认为还可以在某些迭代次数下，就停下来及时进行一批检测，以及时发现是否会有风险

以上是我想到的一些防御方法，主要是从**人为的规定**，增强安全系数，降低有人去破坏模型和**技术的防御**，去不断检测和增强模型的鲁棒性两个大方向入手的。

问题四

对于已经检测出含有恶意后门的模型，你觉得什么样的补救措施或许能够消除或缓解后门攻击对模型的不良影响?

- 检测和隔离后门：如果无法移除后门，则可以考虑使用检测技术来识别后门并将其隔离。这可以通过使用一些技术，如**模型修剪**等来实现。模型修剪可以减小模型的规模，并使其更加简单，以隔离可能存在的后门。
- 进行模型蒸馏：为了减少后门攻击的影响，可以采取一些技术来提高模型的鲁棒性，如模型蒸馏等。设置**大量正确的数据**来进行蒸馏训练，它不仅可以实现模型规模的缩减，同时使模型更加鲁棒，从而降低后门攻击的不良影响。
- 模型再训练：通过在模型中引入一些随机性来增加模型的不确定性，从而降低后门攻击的效果。这可以通过使用Dropout等技术来实现；将已经包含后门的模型重新进行训练，以消除其中的恶意后门；扩大原始的数据集，利用现有模型作为初始参数，再次训练，耗时会降低，并且可以调整参数

总结而言，我能想到的方法较少，主要都是要对模型做技术上的修改的，要么就是增加新的层数，让新的层起到辨别的作用；要么就是把模型重新训练一下，整体性的调整，通过新的好数据集，因为有了初始化参数后调整是较快的；要么就是训练一个全新的模型，利用知识蒸馏的方法去实现了。