Homework

干净标签数据投毒和脏标签数据投毒各有什么优缺点？前者一定比后者好吗？

干净标签数据投毒攻击时一种具有使用价值的新型投毒攻击， 中毒图像标签与视觉感官一致， 攻击者通过少量的投毒使选定的目标测试图像被错误分类。 与此同时受害模型具有较高的分辨率。而脏标签攻击就是一种比较经典且老式的， 通过改变训练数据中某一种物体的部分标签使得训练模型出错。脏标签的攻击优点在于比较容易部署， 且对于攻击者来说并没有很多消耗。 它的缺点也十分明显， 如果有人检测输入的数据， 那么它的标签不符的话就很容易看出来。 与此同时， 干净标签的优点在于很难被人为发现， 因为在人为筛选中毒数据的时候可能因为看起来的结果与标签一样， 所以很难区分 ， 导致这种攻击可能很难被人工筛查。 其缺点是部署的花费较大， 并且不一定可以百分之百找到一个小的扰动使得最后模型训练出一个不同的模型结果， 并且他要求的条件比较苛刻， 如经典的特征碰撞法需要白盒攻击。 我认为前者不一定比后者好， 比如如果有对大规模模型进行数据投毒并不要求具体的分辨类别的话， 脏标签数据投毒会更容易部署。

基于k-NN的中毒把数据检测有什么优缺点？这种检测方法对干净标签数据投毒和脏标签数据投毒都有效吗？

k-NN即k临近算法， 如果一个样本在特征空间中的k个最相似（即特征空间中最邻近）的样本中的大多数属于某一个类别， 则该样本大概率也属于这个类别。 这种方式的优点是它的正确率很高， 且不会有大量的良性数据被误认为是中毒数据。 它的缺点是对其中的超参数的选择十分重要， 如果k的选择不对会有很大的影响。 如果k的值较小， 可能导致过拟合， 如果k值较大则会欠拟合。 这种检测方法对干净标签和脏标签的数据投毒都十分有效， 但是只是被污染数据较小的情况下， 如果存在大量的污染数据可能就会有一定的改变。

为了防范数据投毒攻击， 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？

一， 对模型本身而言， 加强模型的鲁棒性是一个通用的防范投毒攻击的方法。比如集成学习——利用多个基模型的训练结果来产生一个总的结果。 再比如加入生成的数据， 比如mixup方法通过生成一系列离散的数据来达到让数据联系化， 从而你和实际的样本分布情况。 再或者本身针对模型加入一些污染数据来调整它的训练参数。 最后其实可以让训练的过程中让人工筛查每个数据的输入和模型训练结果以达到避免污染数据的加入。