南 开 大 学

本 科 生 毕 业 论 文 （设 计）

中文题目： **基于投票机制的神经网络后门样本检测**

外文题目： **Neural network backdoor sample detection based on voting mechanism**

学　　号： 1811464

姓　　名： **郑佶**

年　　级： 2018 **级**

学　　院： **网络空间安全学院**

系　　别： **信息安全**

专　　业： **信息安全**

完成日期： 2022 **年** 5 **月**

指导教师： **张玉　副教授**

关于南开大学本科生毕业论文（设计）的声明

本人郑重声明：所呈交的学位论文，是本人在指导教师指导下，进行研究工作所取得的成果。除文中已经注明引用的内容外，本学位论文的研究成果不包含任何他人创作的、已公开发表或没有公开发表的作品内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明。本学位论文原创性声明的法律责任由本人承担。

学位论文作者签名：

年　　月　　日

本人声明：该学位论文是本人指导学生完成的研究成果，已经审阅过论文的全部内容，并能够保证题目、关键词、摘要部分中英文内容的一致性和准确性。

学位论文指导教师签名：

年　　月　　日

摘　要

在人工神经网络模型相关的研究中, 有关神经网络的训练和应用中存在的问题一直是研究者探讨的重要方向. 而在神经网络相关的研究中, 对卷积神经网络的研究是一个被广泛运用且应用价值很高的研究领域. 本文通过使用差异化训练的模型集合对特定测试集进行投票式的预测, 并通过分析预测差异的特征, 来判断测试集中恶意样本的存在.

GitHub repo: https://github.com/DonquixoteGarry/test

关键词：投票机制; 后门检测

Abstract

In the research of ANN's models, the puzzles about training and application of them are always a main direction for researchers. And in the research of DNN, CNN is a widely applied direction with high value. In this essay, via the predicting the results for test dataset by a model set which is trained with different training dataset, we can analysis the features about them and distinguish differences of malicious samples and normal samples.

Key Words: voting mechanism; backdoor detection

目录

摘要 …… I
Abstract …… II
目录 …… III
第一章 绪论 …… 1
第一节 人工神经网络与神经活动研究 …… 1
第二节 人工神经网络发展和深度学习 …… 4
第三节 全连接网络与卷积神经网络 …… 5
第四节 卷积神经网络与图像识别 …… 7
第二章 神经网络应用的安全问题 …… 8
第一节 神经网络安全问题的常见场景 …… 8
第二节 针对人工神经网络的常见攻击与防御分析手段 …… 9
2.2.1 常见的攻击手段…… 9
2.2.2 具体的先进防御分析手段 …… 10
第三节 本文的观点 …… 15
第三章 卷积神经网络的投票式模型…… 16
第一节 数据集的选取…… 16
第二节 卷积神经网络参数的选取和调整 …… 16
第三节 样本集的投毒预处理 …… 18
第四节 模型的差异化训练 …… 21
第五节 投票机制分析与测试原理 …… 22
第六节 投票机制测试结果的评估分析 …… 23
第四章 总结与展望 …… 25
第一节 对实验设计的总结 …… 25
第二节 实验的不足与展望 …… 25
参考文献 …… 27
致 谢 …… XXIX

个人简历 …… XXX

第一章　绪论

人工智能技术, 本质上是寻求以某种形式和程度上的自动化, 以求在特定方面替代在各种生产与生活领域对人力资源以及人力控制的需求. 其作为目前国际学界最为关注的研究领域之一, 其对社会发展的影响力和推动力难以估量.

而人工智能技术所需要的具有应用价值的人工智能系统, 必须能够在一定程度上模拟人的控制能力. 而模拟人的控制能力, 在理论上需要机器学习的手段和载体以及对相关能力特征的表述. 而 ANN (artificial neural network , 人工神经网络) 与深度学习, 正是被广泛运用的人工智能的载体与学习手段.

在计算机科学领域, 由于在大量人工智能需求场景下都具有的通用性和有效性, 使得如今在图像以及语言的识别预测和控制 [1], 以至更广泛的医学以及生物研究乃至社会经济等领域, ANN 和深度学习技术都是最具价值的研究方向之一.

第一节　人工神经网络与神经活动研究

人工神经网络技术领域具有很深刻的多学科领域交叉的历史背景, 可以追溯到医学和生物研究领域对人类神经活动的研究模拟. 而正是对神经系统和神经结构特性的不断模拟和数理化抽象, 促进着人工神经网络形式和功能上的不断复杂化和精确化, 这是一个不断进步的过程.

奥地利医生 Franz Joseph Gall 通过对人类神经组织切片的微观分析, 得出了人类神经活动依赖于人体脑部的功能的论断, 解释了人类神经功能的物质基础. 在一定程度, 这与神经网络所需求的分析能力, 某种程度上依赖于人工神经网络的数理模型结构, 逻辑关系上十分相似.

意大利细胞学家 Camillo Golgi 与西班牙神经组织学家 Santiago Ramón y Cajal 通过使用 Golgi 染色法等更精细的微观分析手段, 确认了人类神经组织中神经元功能和结构的独立性. 而神经元结构与功能上的独立性的科学发现, 也为此后人工神经网络中仿神经元的计算单元设计提供了借鉴.

最终在 1943 年, 基于 Franz Joseph Gall , Camillo Golgi 和 Santiago Ramón y Cajal 对人类神经功能运行模式的一系列深入研究, Warren McCulloch 和 Walter Pitts 首次提出借鉴已知神经细胞运行机制的数学模型 M-P 模型. M-P 模型见图 1.1 .

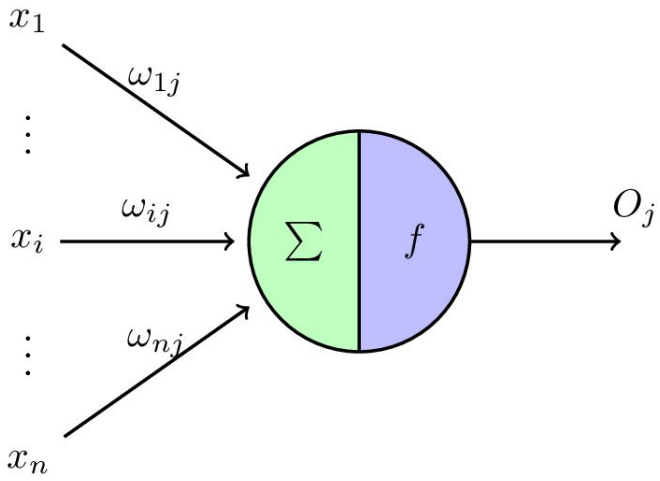

图 1.1 M-P 模型

M-P 模型作为基于简单的函数运算和阈值逻辑来识别输入的二分类器人工神经网络, 是最简单的人工神经网络的架构之一, 首次在数学和计算模型领域引入仿生神经网络的思想, 开辟了人工神经网络研究这个新的计算机科学领域.M-P 模型的提出, 证明仿神经网络的数学模型在一定程度上可以实现逻辑和算术函数映射的功能. 而随后的一系列神经功能运行机制在数理上的抽象和在数学模型上的引入不断强化着人工神经网络模拟复杂映射能力.

而 20 世纪 40 年代末的 Hebb 和 O. Donald 通过在数学模型中引入对神经元的激活机制的抽象, 提出了用以调整其数学模型参数的 Hebb 学习规则, 以模拟神经元的差异性激发对生物神经元间连接强度的影响 [2]. 1957 年, Cornell 航空实验室的 Frank Rosenblatt 提出的模式识别算法感知机神经网络, 即 Perceptron 神经网络, 通过简单四则运算实现了结构简单的双层网络, 并且数理化表述了感知机中尚无法实现的异或回路机制 [3]. Perceptron 神经网络引发了学界对神经网络结构和相关学习算法的广泛深入研究. 其后, Stanford 大学教授 Bernard Widrow 和学生 Ted Hoff 也在 Perceptron 模型做出了基于 ALN (Adaptive Linear Neuron , 适

应性线性神经元) 的改进型的 Adaline 网络 [4]. Perceptron 模型和 Adaline 模型见图 1.2 .

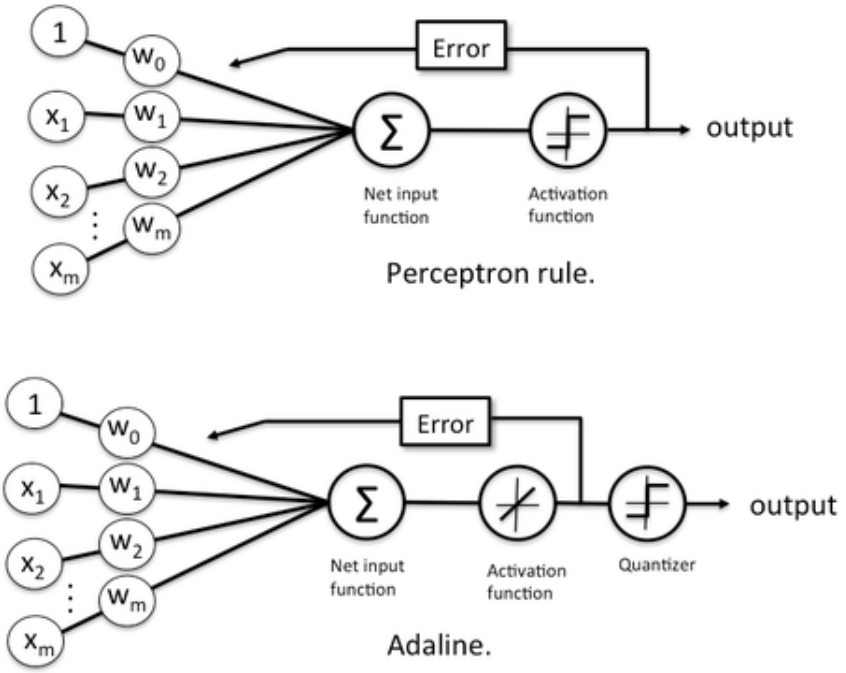

图 1.2 Perceptron 模型以及其改进模型 Adaline 网络

但是上述的 M-P 模型与 Perceptron 模型及其改进型作为早期神经网络模型的代表, 在 1969 年被 Marvin Minsky 和 Seymour Papert 证明其功能上的有限性, 尤其是无法实现 Frank Rosenblatt 所提出的异或逻辑机制, 这一度成为了该领域的研究瓶颈.[5]

随后, Paul Werbos 提出的误差反向传播机制, 即 BP (Back-propagation, 反向传播) 算法的提出, 使得异或逻辑回路的实现在理论上出现了可能, 但是在网络神经元结构上的限制使得 BP 算法仍难以得到有效利用.

John Hopfield 与 Hinton, G. E. 和 Sejnowski, T. J. 分别在多层人工神经网络领域中引入全互联机制和隐单元结构, 使得神经网络领域再次进入蓬勃发展时期. 而 David E. Rumelhart, Geoffrey E. Hinton 和 Ronald J. Williams 提出的非线性 sigmod 函数神经元与误差反向传播算法即 BP 算法的结合, 解决了异或逻辑回路

问题. 1982 年 J.J.Hopfield 提出的 Hopfield 网络在人工神经网络训练领域引入物理学动力学概念, 在在网络输出滞后影响下证明了带有动态非线性反馈的模型训练可以使得模型功能特征状态达到稳态 [6], 这为 BP 算法的效果提供了理论支持. 另外,1997 年 Sepp Hochreiter 和 Jürgen Schmidhuber 提出对人类神经活动中记忆的遗忘机制进行抽象的 LSTM 机制即长短期记忆机制 [7], 这也进一步加强了人工神经网络在运行原理上对神经功能的仿真程度.

像这样的大量的有效的对神经功能模块进行数理化仿真的功能单元的引入, 并且使得相对复杂的非纯线性多层神经网络, 成为人工神经网络结构的重要组成形式, 使得人工神经网络的训练和广泛应用更具有可行性.

第二节　人工神经网络发展和深度学习

深度学习作为机器学习的重要形式, 其发展和人工神经网络的架构的发展也存在一定的同步性, 两者都在一定的层面上存在对人工神经功能的模拟借鉴.

人工神经网络是在生物学研究的基础上, 通过多学科交叉领域学界的的探索, 最终衍生出的计算机科学研究领域. 按照机器学习以及认知科学领域目前普遍认同的定义, 人工神经网络是一种可以根据外部信息进行自适应的仿生数学或计算模型, 这明显是对生物神经系统学习能力的数理化抽象和应用.

而深度学习同样存在对生物神经机理的抽象.20 世纪到如今不断发展的脑科学技术研究, 除了在组织和细胞层面进行结构和功能分析, 其在大脑各分区的功能判断也对人工智能技术发展有所助益. 大脑新皮层感知能力的发现, 成为其中典型的样例.

研究表明, 大脑新皮层作为哺乳动物很多感知能力的物质基础, 其结构上不依赖于对外部刺激的非结构化预处理, 而是将时间上连续的外部刺激信息通过模型结构层次式传递处理 [8]. 经过大量相关的实验, 研究者发现在长时间针对视觉样本的特定训练下, 训练目标能力的图像化边界不断地从粗糙变得精确化 [9].

学界认为机器学习在深度学习全面发展前, 称为浅层学习的学习形式, 是对已知神经功能运行机制的初步抽象 [10, 11]. 在对浅层学习技术为主的时代, 对在高维样本特征学习的过度困难无能为力, 即 Richard Bellman 所称的的” 维度灾

难”. 但是神经结构感受能力在长时间训练中判别能力边界的不断精确化, 给予了深度学习中的特征学习有意义的借鉴. 大脑性皮层对于与数据在感知模块中长时间的层次性传播对学习能力的实现, 在某种程度上, 依赖于对高维样本特征的降维处理 [12], 这在大幅降低神经网络输入数据量的同时能够得到较好的特征学习效果 [13].

在此基础上, 深度学习在人工神经网络中的体现, 可以归纳为以样本处理的手段学习其中的某些复杂的分析特性与分布规律, 使得样本在经过模型处理的过程中, 不断使分布式数据特征精确化, 而这些特征表达则是目标分析处理能力的数理化表述.

第三节　全连接网络与卷积神经网络

基本的非纯线性多层神经网络, 在研究的早期被认为在应用领域具有巨大的价值. 而且在 1989 年, 通过大量针对包含隐单元和非线性单元结构的多层神经网络中 BP 算法性能的探究, 最终在理论上证明了在神经网络层数和隐藏层数足够的情况下, 基本非纯线性多层连续前馈神经网络可以任意程度逼近任意的映射.

但是, 实际上应用这种思想的全连接神经网络在实际训练和应用中效果并不理想. 虽然理论上全连接神经网络能够拟合任意的映射, 但实际上过度复杂的神经网络结构会造成得到目标分析处理能力和参数良性收敛的失败, 使得基于海量数据集的有效深度学习变得困难. 因此, 为增强人工神经网络在深度学习训练上的易行性, 必须在其数理结构上复杂性和实际应用上的有效性中做出取舍.

日本工程师 Kunihiko Fukushima 提出了 Neocognitron 网络, 并在其中引入了” 卷积” 和” 池化” 等不包含在传统非纯线性多层神经网络中的功能概念 [14], 这使得神经网络在组织结构上出现进一步的复杂化.1989 年 Yann LeCun 提出的 LeNet 系列卷积神经网络 [15], 将上述的新概念引入模型应用领域. 通过对基本非纯线性多层神经网络的针对性改进而产生的卷积神经网络, 具有与传统的全连接网络不同的神经网络架构, 很好的解决了全连接网络在实际应用中一部分缺陷.

这是因为, 在卷积神经网络在传统的全连接神经网络结构之外, 还引入了以下的结构 [16]:

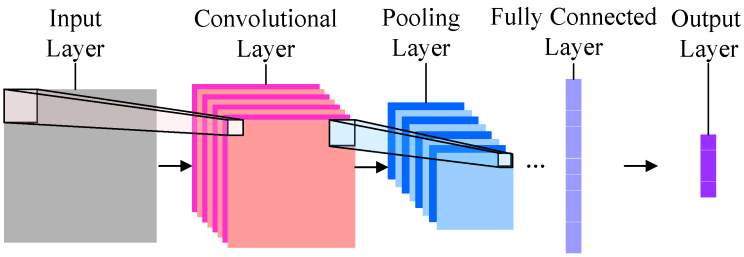

图 1.3 卷积神经网络层级结构

1. 激活层 RELU : 应用非线性激励函数的非线性层, 具有对线性映射性能不足进行补充, 并使输出控制在一定范围内的功能, 作为卷积层和输出层的一部分.
2. 池化层 POOLING : 又称子采样层或汇聚层, 具有在不同深度上子采样 (或译欠采样, Subsampling), 汇聚特征并降低特征的维度, 保留特征提取的高稳定性、显著性和平移不变性 [17], 防止过拟合的功能.
3. 输入层 INPUT : 预处理多维输入, 具有将将输入数据去均值和归一化, 再在各个维度上降维形成若干不相关的特征轴功能的神经元层.
4. 卷积层 CONV : 一种基于在各神经元多维感受域下的局部感知效应的而实现参数共用的复杂计算单元层.
5. 输出层 OUTPUT : 神经网络的最后一层, 由线性层和具有概率分布映射功能的 $softmax$ 函数组成, 视为多分类器.

典型卷积神经网络的基本结构见图 1.3 .

在这些新的结构中, 输入层实现了复杂多维数据在进入网络前的规范化处理, 复数的卷积层的并用能够更加充分的利用输入的多维特征; 而子采样层的欠采样功能, 则能够利用卷积神经网络在连续批量处理数据时的平移不变性等特性, 以及神经元的局部感知效应, 实现权重共享并抛弃冗余的多维特征参数 [18], 并在一定程度上避免过拟合的情况; 而输出层作为标签预测模型的重要构件具有求取各目标标签概率分布的功能.

而总体上来说, 由于卷积神经网络存在模型参数共用和多维特征采集的机制, 导致实际上模型的参数量更精简. 由于卷积神经网络在模型权重参数上的有效精简, 使得其无论是在深度学习训练难度还是实际使用场景中的有效性都更高

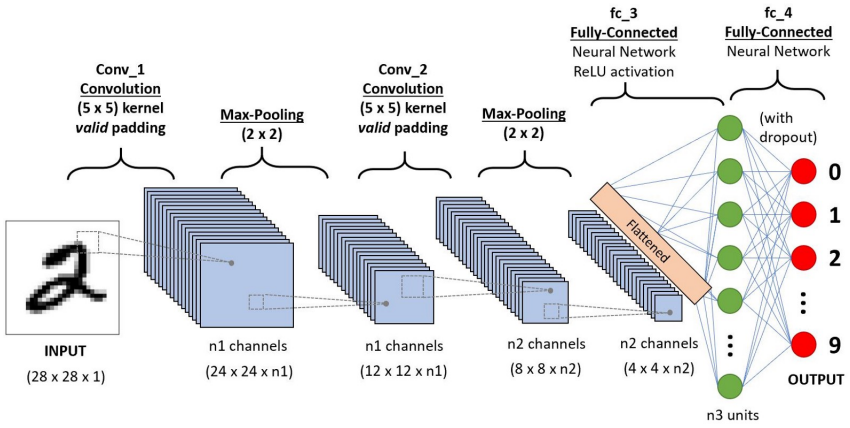

图 1.4 卷积神经网络在图像识别上的应用

[19].

第四节 卷积神经网络与图像识别

神经网络在图像识别领域的应用, 是推动人工神经研究领域进步的现实动力之一. 在诸多的人工神经网络模型概念中, 卷积神经网络是最有应用价值和研究价值的领域之一. 卷积神经元相比全连接的人工神经网络, 更加具有广泛应用的潜力.

1998 年, Yann LeCun 与其他共同研究者, 在自己提出的 LeNet 卷积神经网络模型的基础上提出了改进的 CNN LeNet-5 人工神经网络, 用以对美国的支票等文书上的手写数字进行精确识别 [20] 并证明其应用价值. 这种人工神经网络在社会经济领域的直接应用, 极大地促进了人工智能领域学界对神经网络实际运行性能追求和相关算法和结构的改进, 并使得标准化的人工神经网络图像识别成为衡量神经网络性能的重要指标之一.

卷积神经网络相对于全连接的神经网络, 能够有效避免全连接网络对多维输入向量化造成的信息损失, 同时也避免了实际应用中全连接网络中大量的冗余参数造成训练困难和过拟合现象. 因此, 在图像识别领域的应用上, 卷积神经网络的确相对其他更原始形式的人工神经网络更加具有竞争力和实用性.

LeNet 卷积神经网络在 MNIST 手写数据集上的应用见图 1.4 .

第二章　神经网络应用的安全问题

因为人工神经网络技术在社会各领域的广泛运用, 使得人工神经网络的应用项目本身变成了具有重大政治、经济乃至文化价值的影响目标. 因此, 围绕人工神经网络在应用上的安全性, 对相关攻击手段和防御方式开始变得越来越被重视.

第一节　神经网络安全问题的常见场景

目前, 围绕人工神经网络实际应用的各类攻击手段并不统一. 若不在应用场景中没有第三方情形下考虑安全问题的话, 则可以通过第三方在人工神经网络应用场景中的不同参与方式, 来区分不同类别安全风险的特点. 现代社会的人工神经网络应用中, 第三方的人工神经网络计算平台以及第三方的数据集和模型, 都是潜在的第三方参与的有安全隐患的应用场景.

根据这些这些常见的风险形式, 我们可以划分出第三方平台、第三方数据、第三方模型三个主要场景.

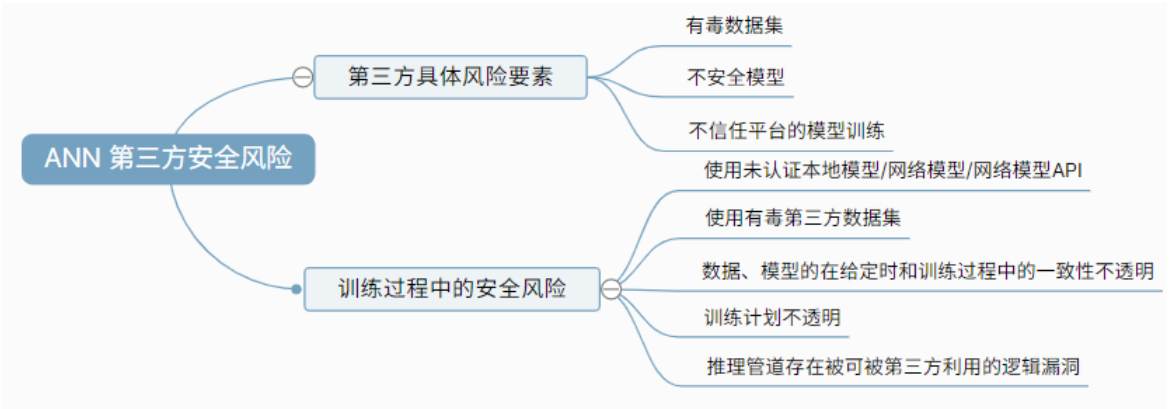

图 2.1　人工神经网络应用中的风险

在第三方平台控制下进行训练, 存在模型和数据被篡改的风险. 虽然第三方平台可能实现对运行参数的透明, 但仍然不能排除训练过程中暗中修改模型或训练过程计划, 以及修改用户方提供的良性数据集插入有毒数据等恶意行为的存在. 对这样的场景尚没有能完全消除恶意风险的手段, 一般可以通过在良性环境下重复训练以调整恶意修改模型造成的效果.

在可能恶意的第三方数据集的影响下, 除了通过某些方式清除有毒数据外也没有根本的解决方式. 但是这种情形下的安全风险仅限于有毒数据, 而无法对模型结构、训练过程计划、推理管道造成影响.

有恶意风险的第三方模型, 一般在应用场景中通过互联网和源码非透明公开的 API 引入. 这类型的安全风险在几类场景中是最大的, 因为有毒的模型可以污染模型无关的推理管道外几乎所有的处理过程. 针对这种安全风险, 需要在人工神经网络输入数据的预处理阶段或是推理管道的功能上做出有效的防范 [21].

第二节 针对人工神经网络的常见攻击与防御分析手段

2.2.1 常见的攻击手段

除了针对人工神经网络的攻击手段不统一之外, 攻击所期望达成的形式和目的同样具有差异性. 按照类似于计算机病毒威胁的划分, 我们可以将针对神经网络的攻击划分为非指向性的功能失效与有指向性和目的性的功能变化诱导.

功能失效的主要体现是人工神经网络功能的普遍降低或失效, 在常见的分类器神经网络模型上, 具体表现可以是对各类测试样本分析有效率的全面降低. 在这方面典型的攻击手段是第三方的普遍数据投毒, 因为普遍数据投毒会造成神经网络决策边界错误转移和混淆.

而功能变化诱导则是需要藉由攻击, 影响人工神经网络实现新的特定功能, 一般不会对网络功能做出彻底破坏. 在这方面的典型手段则是后门植入, 也是本文所关注的重要攻击手段. 后门植入可以在分类器上实现对特定类指向性的分析功能破坏.

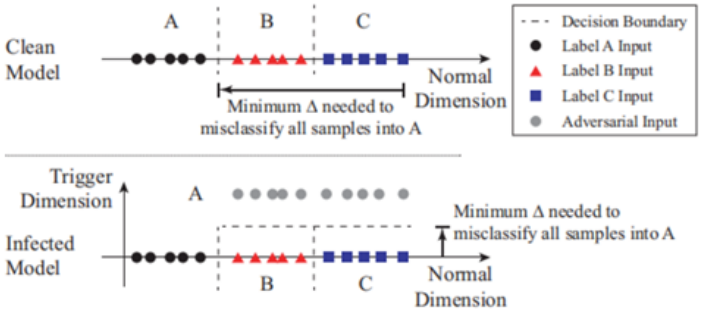

图 2.2 神经清理法假设中触发器对分类的对比影响

2.2.2 具体的先进防御分析手段

近些年来, 人工神经网络的攻击技术的发展促使着防御分析技术同步发展, 以 NC 法, AC 法, SentiNet 法, STRIP 法为代表的等多种防御分析手段体现了重要的研究价值. 如下我们对这些防御分析方法进行原理和运用上的分析.

1. NC 法 (Neural Cleanse , 神经清理): 一类基于分类特征抽象分解的针对后门植入的防御手段.

   若将样本的特征区分抽象视为在模型分类维度上的区分和后门触发器维度上的区分的复合, 那么此类防御基于两个假设. 假设一, 带后门的模型对应的触发器比正常模型形成的触发器相关决策边界更小. 假设二, 在感染模型中能引起误分类的沿触发器方向的最小变化量小于正常模型 [22].

   后门分析过程中, 需要通过反向构造将输入预处理为有相关系数的带触发器的输入, 并通过这些带触发器的输入求取结果的 L1 范数来判断触发器的大小. 当对应的异常指标过大时则认定需要神经清理. 在清理过程中, 根据神经元的激活值选择性裁剪神经元, 移除激活值畸高的节点分支.

   这种防御方式同时具有具体性能和理论上的缺陷. 是有效的清理过程虽会大幅攻击成功率, 但同时会小幅降低对良性样本的分析准确度. 而理论上, 神经清理无法也捕获不依赖优势属性的潜在触发器.

2. AC 法 (Activactive Cluster , 聚类激活分析法): 一类基于神经元激活情况

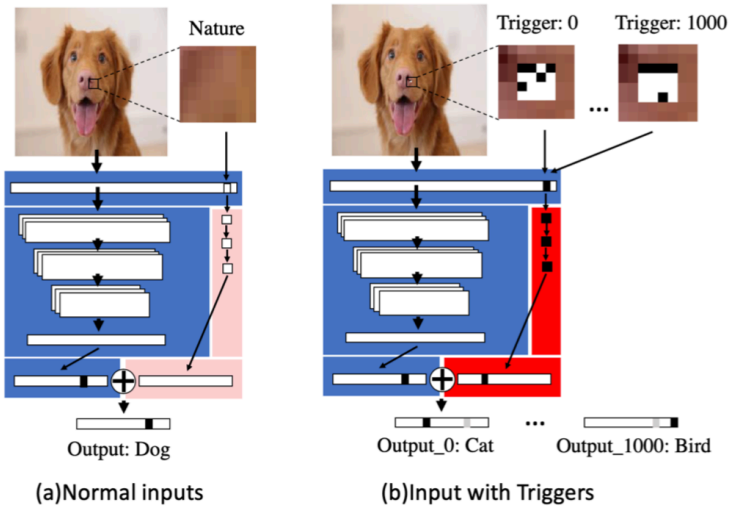

图 2.3 激活聚类法假设中触发器相关类和正常类识别过程的差异

在降维聚类分析后特征的分析手段.

若将后门样本识别为指定类的过程, 视为污染特征对高优先级指定类触发器的触发以及源类特征对正确分类的触发的叠加. 那么由于正常样本不存在对触发器的触发, 两类样本在神经网络中的神经元激活状态也不同 [23].

在操作上, 分析目标数据集在经过神经网络处理过程中神经元的激活情况分布, 需要使用降维和聚类的手段, 将神经网络最后一层输出降维聚类得到分布图.

目标分布图是二维平面上的点的集合, 划分为若干个集中区域称为簇. 一般中毒样本所集中的簇规模相对较小, 而代表正常样本的各个簇的规模往往接近. 当出现规模显然过小的簇时, 一般可以认定数据集合模型受到了污染.

3. SentiNet 法: 一类基于图像多分类器的触发器的触发区域限制的分析手段. SentiNet 法的不仅可以对模型的后门进行分析, 还可以分析甚至定位触发器所在的区域.

所谓的 SentiNet 分析法, 需要选定建议类、划分掩盖区域、选定区域遮掩方式. 其中选定的特定建议类作为遮掩行为调整的对象, 遮掩行为

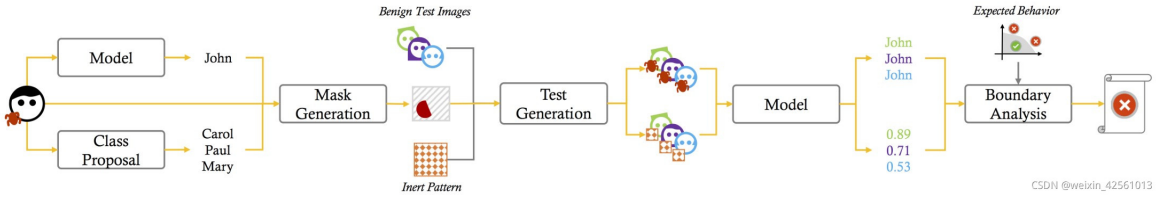

图 2.4 SentiNet 法的分析步骤

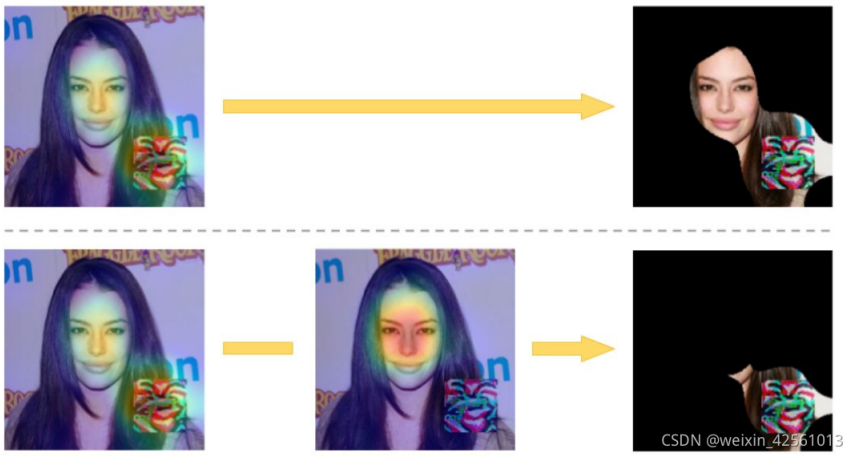

图 2.5 用 SentiNet 法在不同建议类下确定掩盖区域和限定触发器区域

则是通过惰性触发器使得特定建议类的分类能力受到一定程度的限制. 而划分掩盖的区域, 则是在图像中选取触发器可能候选的若干相关区域. 由于触发器所在的候选区域的输入一般在卷积网络运行时较为重要和活跃, 因此 Grad-CAM 热力图算法等手段可以筛选出这些高重要度区域. 而且, 也可以通过不同建议类下求得的掩盖区域的共同特性寻找到触发器所在的区域.

在分析时, 需要在并行的测试中各选定不同的单一建议类, 确定各种建议类下的分类结果的平均置信度和称为愚弄计数的错误分类数并绘制为二维点图. 而测试的输入需要随机噪声和选定污染样本作为对照, 因为两者在二维点图上的拟合特性明显不同.

如上, 根据 SentiNet 分析法提出者 Edward Chou 等人的分析, 二维点图在在随机噪声与对抗性污染样本上的表现明显不同, 随机噪声的二维

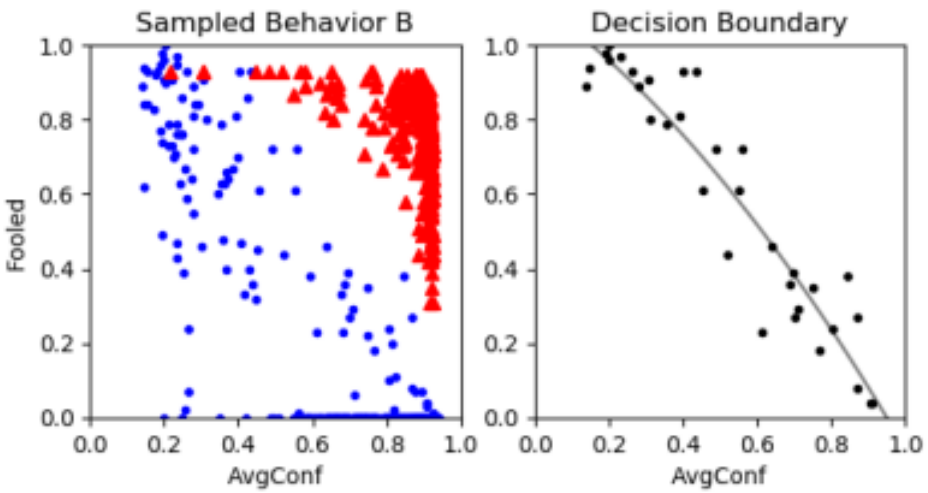

图 2.6 二维点图在在随机噪声与对抗性污染样本上的表现

点图一般可以拟合为直线, 而污染样本在图像上的样本点则一般同时具有高愚弄计数和低置信度 [24].

4. STRIP 法: 一类基于熵分析的对后门植入的分析手段, 这类方法假设带触发器或受污染的输入的结果熵较低.

   在操作中, STRIP 法的后门分析, 基于对任一自然被污染的输入进行特定方式均匀混淆处理而衍生的系列扰乱输入, 在通过带触发器的模型后, 其结果相对低熵的假设. 但这种防御方式虽然可以同时对样本输入和模型本身特性进行分析, 但对未被植入后门的模型无效, 也同时可能会因为扰乱输入对触发器的误触发导致熵畸高 [25].

2021 年, 在 Di Tang 等人针对后门污染分析的论文研究中, 指出了如上的几类防御分析手段缺乏对样本特征的统计化和解析化分析手段的缺点, 提出了统计污染分析法即 SCAn 法 [26].SCAn 法对样本特征的处理手段相对更加注重统计特性.

SCAn 法将样本特征 x 通过映射 R 变换为向量 r, 并进行向量分解, 分解为代表单一类 t 特性的恒常分量 μ_t 和服从高维特定分布的变异分量 ε 之和. 即如下公式:

$$r = R(x) = \mu_t + \varepsilon \tag{2.1}$$

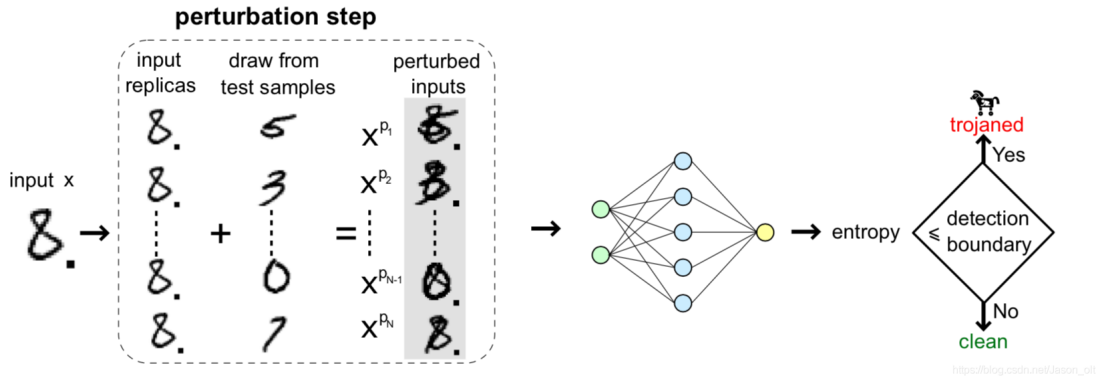

图 2.7 STRIP 法的输入混淆分析

但由于根据实际情况下, 恒常分量可能并不是只代表单一类特性.

Di Tang 等人的分析, 将恒常分量仅代表一个类 t 的假设称为朴素同质假设, 将恒常分量作为源类 t_1 和被攻击类 t_2 特性的线性复合的假设称为双组分分解假设. 双组分分解的假设公式如下:

$$r = R(x) = \sigma * \mu_{t_1} + (1 - \sigma) * \mu_{t_2} + \varepsilon \tag{2.2}$$

根据上述两种假设, 则可以对同一被感染样本点集做出不同的分析.

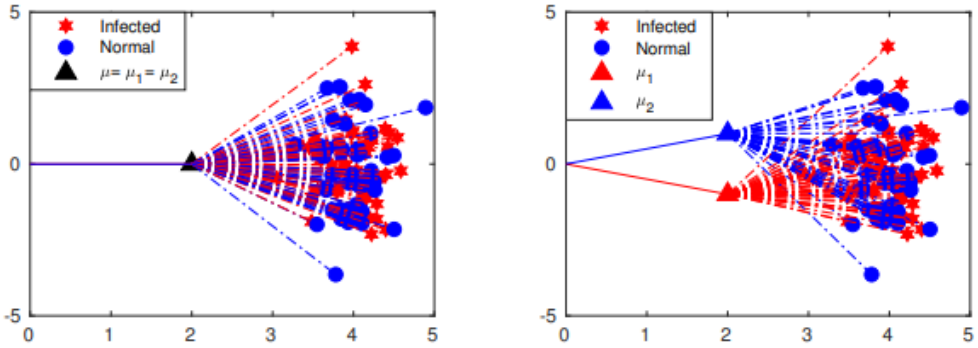

图 2.8 分别按照朴素同质假设 (左) 和双组分分解假设 (右) 对样本点分布的分析

如上, Di Tang 等人的实验证明, 在基于两种假设中, 双组分分解假设明显能

够区分出被攻击类的特征, 应该更适宜被采用为 SCAn 法样本向量分解的参考公式.

运用此公式, 先选定良性样本集条件, 并通过数理统计手段, 估计恒常分量与变异分量分布相关参数的值. 在判断任一不确定样本集时, 将高维分布的参数参考值带入分布类型进行假设检验, 判断参数向量的置信度是否符合假设, 若良性样本集假设不成立, 则表明给定样本是非良性样本集.

由此可见,SCAn 法相对上述其他诸主流方法, 更加依赖统计特性, 具有更高的细粒度. 因此, SCAn 法能够更加精确的区分有毒样本和一般样本, 甚至可以针对不典型的、不规律的分布方式的触发器进行捕获, 在应用上具有相当高的普遍性.

第三节 本文的观点

上述的诸多主流的人工神经网络防御分析手段, 在分析所依靠的假设理论上都有可借鉴的方面. 首先, SCAn 法的双成分分解法, 表明对良性样本集和污染样本集可以依赖统计特征的区分. 而 STRIP 法对人工神经网络输出结果的熵分析, 一定程度上也可作借鉴, 甚至可以推广到模型间的层级上.

在一定程度上, 由于有效神经网络训练规模的庞大性与结构的复杂性, 学界目前对人工神经网络内部权重参数分布特征与神经网络后门间的对应关系仍不能充分解析, 因此本文选择绕开对神经网络本身内部特性的探究, 选择从恶意样本与神经网络后门对神经网络的输出特性的影响进行研究.

为充分研究神经网络的输出特性, 需要从大量神经网络的输出中得到分布的输出规律. 本文意图通过在不同数据下进行差异化训练的同质的卷积神经网络模型在测试集中对特定对象进行分布式的预测, 并通过比较不同模型间的预测差异, 来从目标测试集中提取某些特定样本相对于良性样本的异常特性.

第三章　卷积神经网络的投票式模型

第一节　数据集的选取

MNIST (Mixed National Institute of Standards and Technology database , 美国国家标准与技术研究院混合数据集) 是由美国人口普查局的相关工作人员与一些学生共计 250 余人的手写数字图片经 NIST (National Institute of Standards and Technology , 美国国家标准与技术研究院) 收集整理和数据化的大型手写数字数据库.

在程序中导入的 MNIST 数据集是由向量化的 28x28 的黑白灰度图像和数字 GT 标签 (Ground-True Label, 事实标签) 的样本-标签对组成的集合, 训练集和测试集的规模分别是 60000 对和 10000 对.

灰度值使用 0 到 1 之间的单个浮点数表示,0 表示该像素点全黑而 1 表示该像素点全白. 数字 GT 标签表示为 0 到 9 之间的整数.

MNIST 数据集在人工神经网络的各研究领域中都极为常用, 而且因为其二维图像的性质, 神经网络的后门表现也相对具象化, 因此在本实验中采用该数据集.

第二节　卷积神经网络参数的选取和调整

在卷积神经网络设计领域, LeNet 卷积网络模型系列具有很悠久的发展历史. LeNet 模型作为最早诞生的卷积网络模型之一, 在经过提出者 Yann LeCun 多次的迭代, 得到了开拓性的成果 CNN-LeNet-5 卷积神经网络.

根据上述绪论中的卷积神经网络构件的分析, 可以得知 LeNet-5 神经网络是输入输出层、全连接层、卷积层、子采样层之间的线性连接组合. 而且为适应本次实验对 MNIST 数据集的处理, 需要适当的对标准的 LeNet-5 卷积神经网络的一些参数做出调整. 对卷积神经网络中主要的层级结构的定义和参数调整如下:

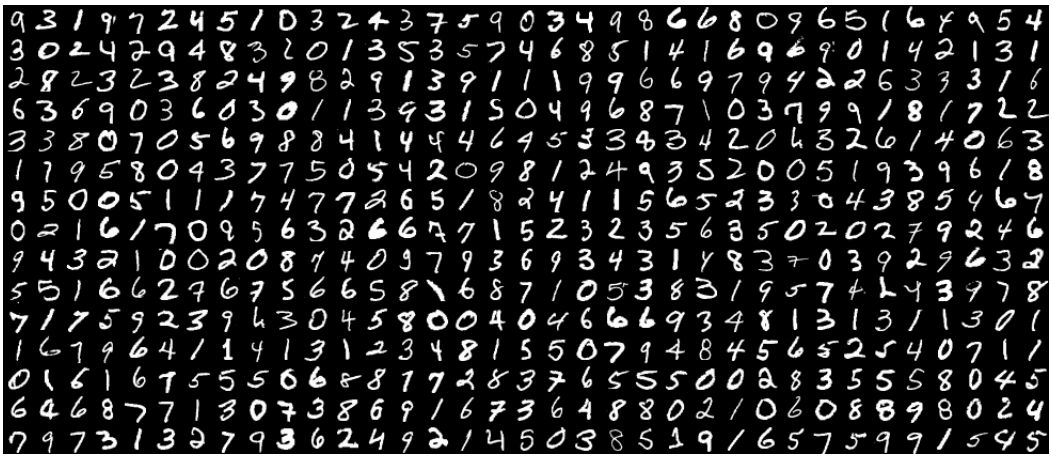

图 3.1　由 MNIST 数据集转换的部分黑白灰度图像

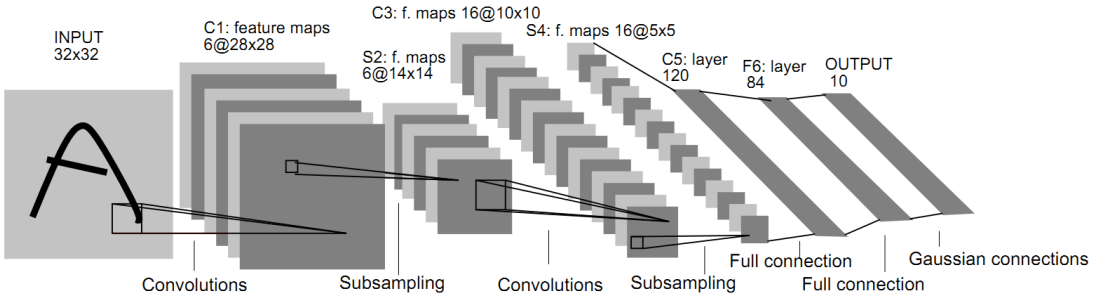

图 3.2　标准的 LeNet-5 卷积神经网络结构

1. 输入层: 输入层的输入张量的格式为 *bs**1*28*28,*bs*(batch size) 即卷积神经网络训练的批量大小
2. 卷积层 *conv*1: 卷积核形状为 5*5, 输入和输出通道数分别为 1 与 10, 采用非扩张卷积法, 并选择卷积核移动步长为 1.
3. 卷积层 *conv*2: 卷积核形状为 5*5, 输入和输出通道数分别为 10 与 20, 采用非扩张卷积法, 卷积核移动步长为 1.
4. 全连接层 $fc1$: 相当于一个线性变换层, 输入张量和输出张量的规模分别是 320 与 50.
5. 全连接层 $fc2$: 线性变换层, 输入张量和输出张量的规模分别是 50 与 10.
6. 输出层: 使用 $log_softmax$ 函数将形状为 10 的一阶张量映射为概率分布向量并对数化.

卷积神经网络的构件中除了基本的输入输出层、卷积层、全连接层外, 还有池化层与激励函数作为其间的连接层, 其实现的下采样和非线性映射功能, 补充了线性映射的不足, 降低了过多冗余特征对训练的不利影响. 此外, 为提升卷积神经网络的效能, 使其并不依赖于局部特征, 需要引入 *Dropout* 层与 *Dropout*2*d* 层, 通过输出值的概率性清零, 实现神经网络层间弱连接功能, 降低过拟合的概率.

对卷积神经网络中主要层级间的连接层级的定义和参数调整如下:

1. *conv*1 层与 *conv*2 层间: 依次为池化核大小为 2 的池化层和 *ReLu* 激励函数的非线性映射层.
2. *conv*2 层与 $fc1$ 层间: 依次为清零概率为 0.5 的 *Dropout*2*d* 层, 池化核大小为 2 的池化层和 *ReLu* 激励函数的非线性映射层. 最后再输入 $fc1$ 层前再进行降维操作降为长度为 320 的向量.
3. $fc1$ 层与 $fc2$ 层间:*ReLu* 激励函数的非线性映射层与清零概率为 0.5 的 *Dropout* 层.

第三节 样本集的投毒预处理

本实验的思想是通过差异化训练的同质化模型对单一样本的分析差异. 因此在本实验中, 我们需要同时训练得到在良性环境下训练的源预训练模型与经过数

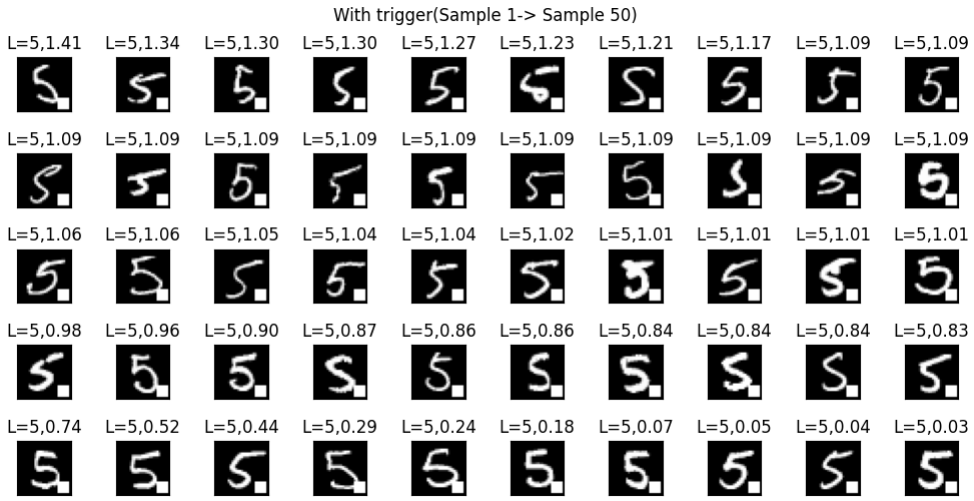

图 3.3 *GT* 标签为 5 的后门样本示例

据投毒毒化训练的毒化预训练模型. 而且, 为了最终测试投票系统的性能, 需要被数据投毒的测试集供其从中判别恶意样本.

因此, 对训练集和测试集的投毒需要能够在模型里植入后门, 并同时在测试时被植入后门模型在恶意样本判别式能体现明显的输出差异, 也就是需要训练集投毒效果的在测试集判别上的高差异性表现. 因此, 我们应当在测试集和训练集中选定不同的单一 GT 源标签对象集合的局部作为预选的样本污染对象.

训练集中选择污染选定某一单一 GT 源标签对象集合的局部, 会使得不同的模型在选取子训练集时就造成差异, 在被污染模型中造成不同的后门植入效果差异. 这样不仅可以在未污染训练的模型对象和被污染的模型间造成训练的差异, 还可以在被污染的模型间形成训练的差异.

为同时提高数据投毒的性能与操作上的易行性, 可以选择典型的二维图像触发器来作为后门恶意样本的特征. 选择典型的二维图像触发器, 不仅有利于具象化的表现后门样本的二维特征, 而且也有利于将人工筛查与投票机制对恶意样本的初步筛查相结合.

实际操作上, 我们可以采取置放方块触发器的方式对特定 *GT* 标签的样本做出污染. 若调用 python 的 pytorch 张量库的对 MNIST 数据集的 dataloader 处理功能, 可以得到每个单一样本的张量化表示. 虽然张量化样本的污染可以简单地用

与张量化的触发器相加表示, 但是由于每个张量元素代表的合法灰度值需要介于 0 与 1, 所以需要规范化的函数不合法的灰度值截断取合法值.

若假设未污染的样本张量化表示为 $V_{original}$, 规范化函数为 $Transform$, 用于计算添加的触发器的张量表示为 $V_{trigger}$, 而被污染后的样本张量化表示为 $V_{perturbe}$, 则污染的过程用以下的公式表示:

$$V_{perturbe} = Transform(V_{original} + V_{trigger}) \tag{3.1}$$

因为正常的卷积神经网络训练与测试中为保证样本调用的随机性, 需要每次都打乱样本集并重新进行批量化分组. 而且,pytorch 张量库的对 MNIST 数据集的处理依赖于官方的文件组织格式. 因此为了程序逻辑上的逻辑简化, 需要以与源数据集文件同样的格式保存被污染的数据集. 由于官方网站提供的数据集是二进制资源格式, 因此上述的数理逻辑可以进一步简化为二进制文件的更改保存. 因此, 只需要对照源文件的组织格式修改特定文件偏移位置上的字符即可.

TRAINING SET LABEL FILE (train-labels-idx1-ubyte):

[offset]	[type]	[value]	[description]
0000	32 bit integer	0×00000801(2049)	magic number (MSB first)
0004	32 bit integer	60000	number of items
0008	unsigned byte	??	label
0009	unsigned byte	??	label
.....			
xxxx	unsigned byte	??	label

The labels values are 0 to 9.

TRAINING SET IMAGE FILE (train-images-idx3-ubyte):

[offset]	[type]	[value]	[description]
0000	32 bit integer	0×00000803(2051)	magic number
0004	32 bit integer	60000	number of images
0008	32 bit integer	28	number of rows
0012	32 bit integer	28	number of columns
0016	unsigned byte	??	pixel
0017	unsigned byte	??	pixel
.....			
xxxx	unsigned byte	??	pixel

Pixels are organized row-wise. Pixel values are 0 to 255. 0 means background (white), 255 means foreground (black).

图 3.4 MNIST 数据集文件的二进制组织规则

第四节 模型的差异化训练

程序差异化训练的目的, 是要通过训练使得两类模型的对后门恶意样本的处理结果差异能在测试中明显体现. 但是参考投票机制与对人工神经网络的 STRIP 熵分析法的相似性, 可知模型本身对后门无关的正确类的低分类成功率, 会使得模型后门植入的影响与模型本身误分类影响相结合, 造成恶意后门样本特征的难以提取.

因此, 为提升训练的效果, 重点在于提高模型本身的分类成功率, 并且使得后门植入的影响更加得以体现. 结合上述对污染环节的分析, 在实际操作中, 我们应该在训练集中, 对设定较高的指定标签样本集污染率, 并设定较小的模型子训练集规模比率, 使得各模型见具有明显但是有一定差异性的后门植入效果.

但在模型的训练中, 还需要考虑数据集外的参数影响, 这包括批量规模和训练轮数, 而且这两个参数的影响并不同质.

在设定批量规模时, 需要顾及本实验设计的特性. 因为本实验在训练集中选取的污染对象是单一 GT 标签样本, 在整个训练集中占比低且分散, 而且是 MNIST 数据集本身是随机打乱的, 所以需要防止数据集的污染” 噪音化”. 所谓的” 噪音化” 值得是由于 LeNet 卷积神经网络等常规神经网络的批量化训练机制, 使得在大批量规模训练中, 少数的分散的样本污染对造成损失计算造成的影响被大幅稀释, 使其如同随机噪音一般成为难以被人工神经网络学习的特征. 为防止训练集中的后门样本” 噪音化”, 我们一般选择降低批量规模大小.

在设定训练轮数时, 主要考虑限制过拟合现象和神经网络损失不收敛的可能性. 过低的训练轮数可能造成人工神经网络对特征学习的不充分和网络损失的不收敛, 而训练轮数过高时同样可能造成过拟合现象的发生.

但同时, 训练时间和空间代价对这两个参数的调节也是重要的考虑因素. 过高的批量规模对 GPU 硬件造成的运算压力过大, 过低的批量规模和过高的训练轮数会造成成倍的多余时间开销.

第五节 投票机制分析与测试原理

本实验最终的性能体现需要靠最终的多模型测试集判别器, 也就是所谓的投票模型实现. 通过投票模型, 最终得到的判别数据是由一系列的概率分布向量组成的矩阵.

设存在预训练洁净模型 M_1 至 M_n 与经污染化训练的预训练模型模型 M_{n+1} 至 M_{2n}, 模型 M_i 对样本 S 的概率分布输出是向量 v_i.v_i 的各元素是由人工神经网络的输出层的 $softmax$ 函数输出的 10 个数字类的概率值, 格式如下:

$$v_i = [p_{0,i}, p_{1,i}, ...p_{9,i}]^T (1 \leq i \leq 2n) \tag{3.2}$$

根据 $softmax$ 函数的性质则有以下关系

$$\sum_{x=0}^{9} p_{x,i} = 1 (1 \leq i \leq 2n) \tag{3.3}$$

而对于样本 S 的判别数据或称判别矩阵 $Matrix_S$ 如下:

$$Matrix_S = [v_1, v_2, ..., v_n, v_{n+1}, ..., v_{2n}] \tag{3.4}$$

根据实验的目的, 对于测试集的任何一个样本, 都需要通过其判别矩阵得到异常指标作为衡量样本相对于后门样本的疑似指数. 因此, 将判别矩阵映射为异常指标的计算公式, 要求能够体现模型间对同一个样本的评判差异. 某种程度上, 模型间对同一个样本的评判差异可以表现为各模型对该样本在各类上概率差异的复合. 因此我们可以选择各模型对该样本在各类上概率标准差的和作为异常指标.

在本实验中, 需要将判别矩阵 $Matrix_S$ 按照类切分成 10 个各模型对该样本在同一类上的概率值组成的判别向量 $v_0{}'$ 到 $v_9{}'$, 如下:

$$Matrix_S = [v_0{}', v_1{}', ..., v_9{}']^T \tag{3.5}$$

然后对任一判别向量 v_x 求元素间的标准差 ST_x, 若将这个函数定义为 std, 则

有:

$$ST_x = std(v_x) = \sqrt{\sum_{t=1}^{2n}(p_{x,t} - \frac{\sum_{t=1}^{2n} p_{x,t}}{2n})^2} \tag{3.6}$$

然后便可得到样本 S 的异常指标 I_S 公式如下:

$$I_S = \sum_{x=0}^{9} ST_x \tag{3.7}$$

在实验中, 设定被定为异常样本的最低异常指标值和测试集最高异常指标值 $I_S{}'$ 间的比率, 或称异常置信比 p_{vaild}. 可得含有 m 个污染样本的规模为 N 的测试样本集漏报率 F 公式为:

$$F = 1 - \frac{\sum_{x=1}^{N} Judge(I_{S_x})}{m} \tag{3.8}$$

若将人工识别样本是否含有触发器的结果抽象为自变量为样本的函数 $Trigger$, 则上式中判断样本是否异常以及是否能通过人工识别为有毒样本的函数 $Judge$ 定义如下:

$$Judge(I_{S_x}) = \begin{cases} 1 & \frac{I_{S_x}}{I_S{}'} \geq p_{vaild} \quad and \quad Trigger(I_{S_x}) = True \\ 0 & otherwise \end{cases} \tag{3.9}$$

在本次实验中, 最终得到的漏报率指标 F, 可以即后门样本未被限定在取信为异常样本的集合中的比率. 漏报率指标作为实验性能的最终指标, 其值越低, 表示实验对后门样本与未污染样本间区别特征提取在测试中表现效果越好, 反之效果越差.

第六节 投票机制测试结果的评估分析

通过基于上述基础原理设定的实验的相关结果, 我们可以对投票机制的相关性能做出评估分析.

在实验中, 我们指定以下参数: 洁净预训练模型数和污染模型数均取值为 n,

模型子训练集规模比率 0.2, 异常置信比 p_{vaild} 取值为 0.75, 训练集批量规模取值为 8, 训练轮数取值为 50, 选定的预训练模型的总分类准确率为 98.1%.

训练集指定类污染率设定为 $P_{perturbe}$, 规模为 N 的测试样本指定为 m 个污染样本. 通过更改参数 $P_{perturbe}$ 和 m 重复进行训练和测试. 经过测试, 得到如下的实验数据.

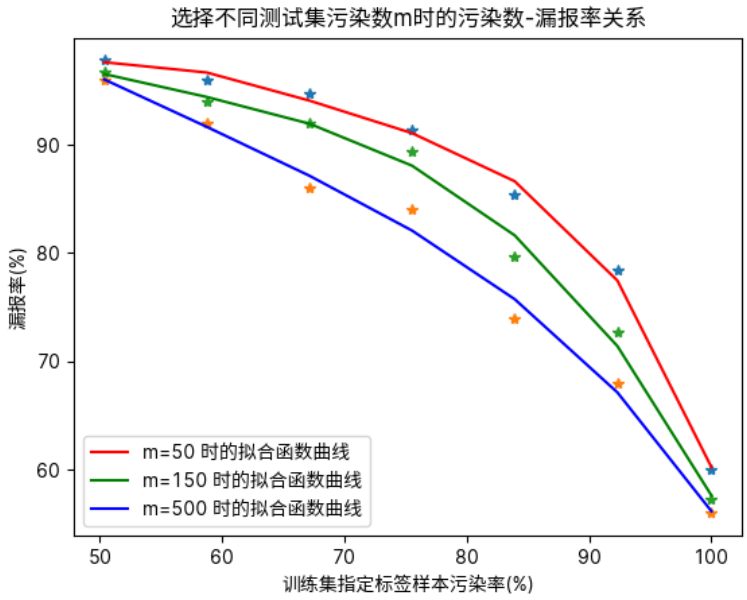

图 3.5 不同的训练集指定类污染率 m 取值下, 对应的平均漏报率 $P_{perturbe}$

根据上面得到的数据可以得知, 在各投票模型的漏报率随着训练集污染率不断下降, 并在接近 100% 时达到 60% 左右的最低值. 而且, 随着测试集污染数的上升, 漏报率也呈现下降的趋势. 也就是说, 在经过对后门植入效果最明显的污染训练后, 后门样本会向着高异常指标的方向富集. 分析上述的数据, 可以结合前面的分析得到几个结论. 其一, 随着训练集污染程度的提升, 后门植入的效果愈发明显, 在投票模型中对异常指标的贡献更大, 有利于投票机制的性能. 其二, 随着测试集污染程度的提升, 恶意样本占测试集的比率升高, 使得一般样本对异常指标判别的影响下降, 同样有利于漏报率的降低.

第四章 总结与展望

第一节 对实验设计的总结

本文通过基于目前在图像识别领域常用的神经网络架构, 通过搭建模型间投票分析模型实现对测试数据集中恶意后门样本的甄别. 本文搭建的模型间投票分析模型, 是基于对人工神经网络防御分析手段中 STRIP 法的分析手段在模型间输出差异上的应用. 本文通过在特定环境下的投票机制性能的测试, 证实了投票机制在一定程度上的可行性. 投票机制相对于人工识别, 能够在一定程度上更高效的从总样本集中区分出后门样本的特征.

本文完成的工作是:

1. 通过对人工神经网络以及深度学习领域相关领域的研究, 了解该领域的发展脉络和研究价值. 通过相关文献的查阅, 人工神经网络技术应用中可能造成的相关安全风险.
2. 学习人工神经网络的层级结构和功能运行特性与原理, 在此基础上深入理解各类人工神经网络常见的防御分析手段的原理与具体手段.
3. 基于对几类常见人工神经网络的分析理念的发展和借鉴, 搭建基于分析差异化训练模型间输出差异的多模型投票机制, 并通过在理论上探讨相关实验参数的调整对实验效果和目的的影响.
4. 对除不适合的参数外的参数重复调整, 重复进行实验得到相关的数据. 验证投票机制在分辨异常后门样本时的性能. 并分析实验调整的相关参数对实验结果的影响.

第二节 实验的不足与展望

因经验和理论水平所限, 本实验在理论上和相关应用测试设计让仍有不足:

1. 本实验选择调整过常用的 CNN-LeNet-5 模型作为投票机制中特征学习和

概率数据分析的主干网络. 但是仍存在相对更优秀的 Alex Net 等网络架构和算法的模型, 能使得其相对于以 LeNet 系列卷积神经网络在图像识别正确率、防止过拟合现象 [27]、后门植入和体现输入差异有更好表现.

2. 由于实验器材性能的限制, 实验限定了网络结构的复杂度, 并调整了相关参数以降低运行的硬件要求, 这可能会影响投票机制的性能体现.
3. 本文在理念上基于对 STRIP 后门分析法与 SCAn 分析法在模型间输出差异上移植, 选定的异常指标实际上作为 STRIP 分析法中熵分析法的借鉴, 其映射计算方式在体现模型分析差异的方面上仍有优化空间.
4. 本文对模型后门防御机制的分析基于典型的后门触发器, 而不基于样本局部特征的复杂触发器与后门机制可能是此类实验分析的未来方向.

参考文献

[1] 郑远攀, 李广阳 and 李晔. 深度学习在图像识别中的应用研究综述. 计算机工程与应用, 2019, 55(12): 17.

[2] Hebb and O. Donald. The Organization of Behavior. 1949.

[3] F. Rosenblatt. The perceptron - a perceiving and recognizing automaton. 1957.

[4] B. Widrow and M. E. Hoff. Associative Storage and Retrieval of Digital Information in Networks of Adaptive ”Neurons”. Biological Prototypes and Synthetic Systems, 1962.

[5] Minsky, Marvin, Papert, et al. Perceptrons : An Introduction to Computational Geometry. 1969.

[6] Hopfield, J. and J. Neural networks and physical systems with emergent collective computational abilities. Proceedings of the National Academy of Sciences, 1982.

[7] S. Hochreiter and J. Schmidhuber. Long Short-Term Memory. Neural Computation, 1997, 9(8): 1735 ∼ 1780.

[8] T. S. Lee and D. Mumford. Hierarchical Bayesian inference in the visual cortex. Journal of the Optical Society of America, 2003, 20(7): 1434 ∼ 48.

[9] T. Serre, G. Kreiman, M. Kouh, et al. A quantitative theory of immediate visual recognition. Progress in Brain Research, 2007, 165(6): 33 ∼ 56.

[10] 陈先昌. 基于卷积神经网络的深度学习算法与应用研究 [PHDTHESIS], 2014.

[11] 胡清华, 张道强 and 张长水. 复杂环境下的机器学习研究专刊前言. 软件学报, 2017, 28(11): 3.

[12] R. Duda, P. Hart and D. Stork. Pattern Classification: Wiley-Interscience. 2000.

[13] 孙志军, 薛磊, 许阳明, et al. 深度学习研究综述. 计算机应用研究, 2012, 29(8): 5.

[14] K. Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biological Cybernetics, 1980, 36(4): 193 ∼ 202.

[15] Y. Lecun. Generalization and Network Design Strategies. In: Connectionism in Perspective, 1989.

[16] 谭庆波. 机器学习与图像处理, 卷积神经网络详解, 2020. https://zhuanlan.zhihu.com/p/47184529.

[17] 张润 and 王永滨. 机器学习及其算法和发展研究.

[18] 常亮, 邓小明, 周明全, et al. 图像理解中的卷积神经网络. 自动化学报, 2016, 42(9): 13.

[19] 周飞燕, 金林鹏 and 董军. 卷积神经网络研究综述. 计算机学报, 2017, 40(6): 23.

[20] Y. Lecun and L. Bottou. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 1998, 86(11): 2278 ∼ 2324.

[21] Y. Li, B. Wu, Y. Jiang, et al. Backdoor Learning: A Survey. 2020.

[22] B. Wang, Y. Yao, S. Shan, et al. Neural Cleanse: Identifying and Mitigating Backdoor Attacks in Neural Networks. In: 2019 IEEE Symposium on Security and Privacy (SP), 2019.

[23] B. Chen, W. Carvalho, N. Baracaldo, et al. Detecting Backdoor Attacks on Deep Neural Networks by Activation Clustering. 2018.

[24] E. Chou, F Tramèr and G. Pellegrino. SentiNet: Detecting Physical Attacks Against Deep Learning Systems. 2018.

[25] Yansong Gao, Chang Xu, Derui Wang, et al. STRIP: A Defence Against Trojan Attacks on Deep Neural Networks. In: 2019.

[26] D. Tang, X. F. Wang, H. Tang, et al. Demon in the Variant: Statistical Analysis of DNNs for Robust Backdoor Contamination Detection. 2019.

[27] 李彦冬, 郝宗波 and 雷航. 卷积神经网络研究综述. 计算机应用, 2016, 36(009): 2508 ∼ 2515.

致　谢

本研究及学位论文是在我的导师张玉副教授的亲切关怀和悉心指导下完成的. 他严肃的科学态度, 严谨的治学精神, 精益求精的工作作风, 深深地感染和激励着我. 从课题的选择到项目的最终完成, 郑老师都始终给予我细心的指导和不懈的支持. 进行毕业设计的数月来, 张副教授不仅在学业上给我以精心指导, 同时还在思想、生活上给我充分的关怀, 在此谨向张先生致以诚挚的谢意和崇高的敬意.

在论文将完成之际, 我感慨颇深心难平静. 论文自开题到完成, 很多位可敬的师长与同学都为我提供了无私的指导和帮助, 请接受我真诚的感谢. 此外, 感谢培养我长大的父母, 恕我情至深而难以表以言辞!

个人简历

基本信息：

姓名： 郑佶

性别： 男

出生日期： 2000 年 12 月 05 日

通信地址： 天津市津南区咸水沽镇海河教育园区同砚路 38 号南开大学津南校区

电话：15067776526

E-mail: 772734603@qq.com

教育背景：

2018.09-2022.07　南开大学　网络空间安全学院　　信息安全　　学士