摘 要

在计算机科学领域中，人工神经网络(Artificial Neural Network，ANN)的因其应用价值巨大，其安全性问题一直受到学界广泛关注。该文从卷积神经网络(Convolutional Neural Networks，CNN)后门分析相关技术入手，通过对基于对网络输出结果的熵分析法，以及基于对样本特征的统计式分析的统计污染分析法等其他相关后门防御分析技术的借鉴，探索从网络输出结果分析和区分恶意后门样本污染特征的手段。文章通过使用经过对同一总训练样本集取同规模不同子集进行训练所得到的模型集合，对特定测试集进行投票式的标签概率分布预测，并通过分析预测差异的统计特征，来从测试集中区分恶意后门样本的特征。

关键词：人工神经网络;卷积神经网络;图像识别;后门检测;统计分析

Abstract

In the field of Computer Science, technology of ANN (Artificial Neural Network) has very high applicative valuation, so its security problem has been widely concerned by the academic circles. Starting with backdoor analysis technology of CNN (Convolutional Neural Networks) , this paper explores the means to analyze and distinguish the pollution features of malicious backdoor samples from the network output results, by referring to STRIP method and Statistical Contamination Analyzer method. This paper uses the CNN model set trained by different subsets of training set with the same scale to predict the label probability distribution of samples from testing set together like voting, and analyzes the statistical features of the prediction differences to distinguish the features of malicious backdoor samples from the testing set.

Key Words: artificial neural network (ANN); Convolutional neural network (CNN); image identification; backdoor detection; statistical analysis

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。 就其应用上的泛用性和概念上的革命性而言，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，必须成为作为控制能力主体的人在一定层面上的有效替代。 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。 而人工神经网络与深度学习，正是被广泛运用的人工智能能力的载体与学习手段。

在计算机科学领域，由于人工智能在大量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，以至更广泛的医学以及生物研究乃至社会经济等领域，人工神经网络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，这可以追溯到医学和生物研究领域对人类神经活动的研究和模拟。 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在形式和功能上不断复杂化，这是一个不断进步的过程。

医生 F. J. 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脑部功能的论断，这解释了人类神经功能的物质基础。 在一定程度上，这与神经网络所需求的分析能力依赖于人工神经网络的数理模型结构的特性，在逻辑关系上十分相似。

随后，细胞学家 C. Golgi 与神经组织学家 S. R. 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计算单元的模型设计提供了借鉴。

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[1]。M-P 模型如图 1.1 所示。

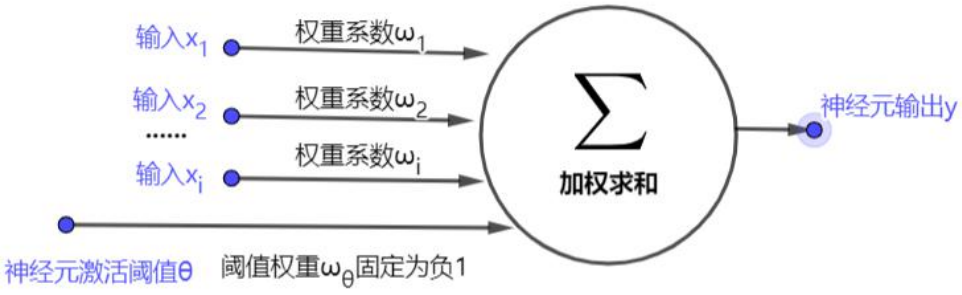

图 1.1 M-P 模型

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，M-P 模型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 D. O. Hebb 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 模型提出了基于自适应线性神经元的作为 Perceptron 改进型的 Adaline 网络[4]。 Perceptron和 Adaline 模型如图 1.2 所示。

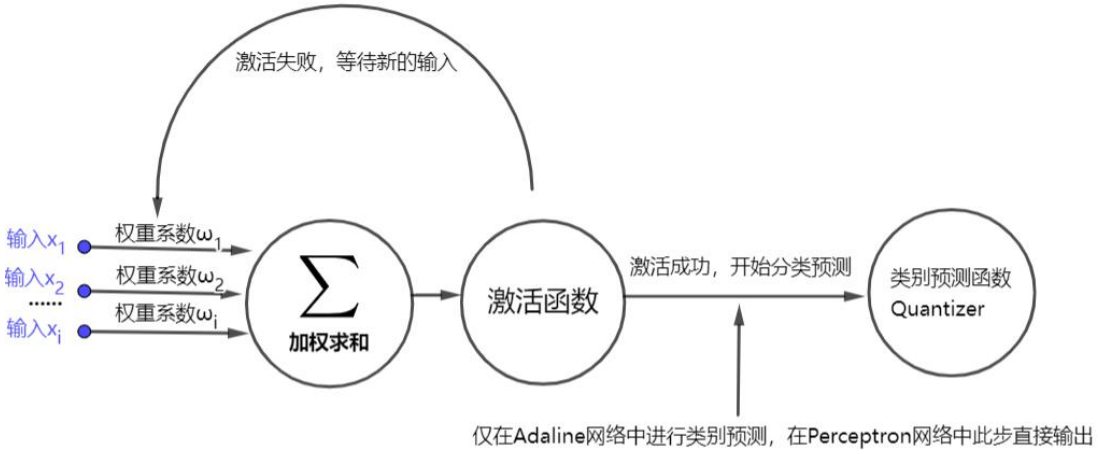

图 1.2 Perceptron 模型以及其改进模型 Adaline 网络

但是上述的 M-P 模型与 Perceptron 及其改进型模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 的 BP (Back-propagation，误差反向传播) 算法，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 函数神经元等新结构与 BP 算法在应用上的结合，解决了异或逻辑在神经网络上的表现难题。 1982 年 J. J. Hopfield 通过提出以其命名的人工神经网络概念模型在人工神经网络训练领域引入物理学动力学概念，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。而 S. Hochreiter 和 J. Schmidhuber 在1997 年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技术

深度学习作为人工神经网络的学习手段，两者都在一定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信号的离散式预处理，而是将时间上连续的外部刺激信号通过模型结构层次式传递处理[8]。 经过大量相关的实验，研究者发现在针对视觉样本的长时间训练下，训练目标能力的可视化边界不断地从粗糙变得精确[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，即发生所谓的维度灾难。 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，也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，最终在理论上被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度水平和实际应用上的效率中做权衡。

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。

因此，为解决基本全连接网络在应用上的问题，卷积神经网络引入了以下的结构:

1. 激活层 :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 :又称子采样层或汇聚层，具有在不同深度进行下采样(或译子采样，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[16]，防止过拟合的功能
3. 输入层 :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 :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 :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* 函数或相似的具有逻辑功能的函数组成

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；而池化层的欠采样功能，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，以及神经元在局部特征域上的感知效应，实现神经网络构件间权重参数的共享并抛弃冗余的多维特征参数[17]，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，使得其特征提取效率和有效性更高[18]。

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 LeNet-5 人工神经网络[19]，用以对美国的支票等文书上的手写数字进行精确识别并证明其应用价值。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与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广泛用于衡量神经网络的性能。

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例如图像等多维输入在向量化降维过程中造成的信息损失，同时也避免了实际应用中全连接网络大量冗余参数造成的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。

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，针对人工神经网络的应用安全性问题，相关的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场景并不统一。若不在没有第三方情形下的应用场景中考虑安全问题的话，可以通过第三方在人工神经网络应用场景中的参与环节差异以划分各类安全风险，具体如图 2.1 所示。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，可以划分出第三方平台、第三方数据、第三方模型等三个主要场景。

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，但仍然不能排除其在训练过程中暗中修改模型或训练过程计划，以及在修改用户方提供的良性数据集插入有毒数据等恶意行为存在的可能。对这样的场景尚没有能完全消除恶意风险的手段，但一般可以通过在良性环境下进行重复训练以平衡恶意修改模型造成的效果[20]。

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运行逻辑上做出有效的防范[20]。

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场景不统一之外，攻击的形式和所期望的目的同样具有差异性。按照类似于计算机病毒威胁的划分，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、目的性的功能转变。

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标签类测试样本预测成功率的全面和大幅降低。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功能决策边界的严重错误转移和混淆。

而功能转变，则是需要藉由攻击影响人工神经网络，以实现新的特定功能，一般不会使原本具有特定功能的神经网络结构性失效，而是造成功能性上的部分混乱。神经网络后门植入是典型的功能转变，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标签类样本指向性的分析功能变化。

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，STRIP 法、基于距离的异常检测法为代表的等多种针对异常污染样本的防御分析手段体现了重要的研究价值。如下是对这些防御分析方法进行原理和运用上的分析。

1. STRIP 法: 一类基于熵分析的对后门植入的分析手段，这类方法假设带相似触发器的输入的输出结果间熵较低。

   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，其系列结果间熵相对低的假设。

   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[21]。STRIP法的典型样本表现和步骤示例如图2.2所示。

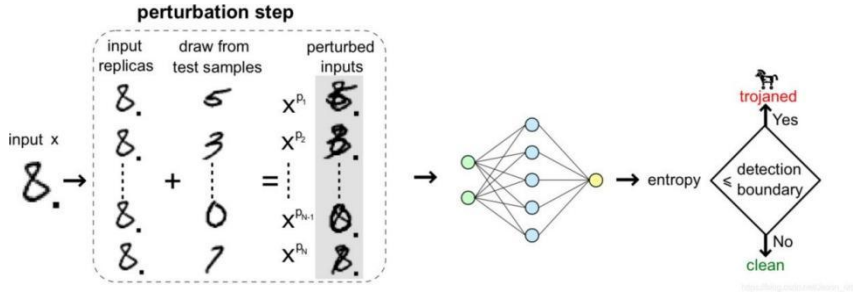

图 2.2 STRIP 法的输入混淆分析[21]

2. 基于距离的异常检测法: 一类基于样本间向量化多维特征间的距离的对后门植入的分析手段，这类方法需要进行未知样本与预设正常样本的比较。

   在操作中，首先在未知父样本集中，通过默认成功且无差错的辨别过程，标定一个完全由正常样本构成的小规模可信子样本集。随后对任何不属于可信子集的样本进行判别时，根据其与可信子集样本间向量化

多维特征间的距离的值与标定的距离阈值作比较，将其划分为正常或异常的样本。

这类防御方法的优势在于，精确分析得到小规模可信样本子集的开销，会明显小于在不划分可信样本子集时，对所有样本间相关联的统计特征的分析[22]。而且在可信子样本集划分精确的情况下，它针对对抗性生样本污染产生很好的防御效果。但是该方法对于距离阈值的调整会影响判别的效果，在参数不合适的情况下该方法的性能可能很低[22]。

但在 2021 年，在 Di Tang 等人在针对后门污染分析的论文研究中，指出当前的防御分析手段缺乏对样本独立分布特征的统计和精确分析手段的缺点，并提出了统计污染分析法即 SCAn 法[23]。

SCAn 法的初步假设中，将样本特征 x 通过映射R 变换为向量 r，并进行向量分解。最终分解为代表单一类 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:

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

但恒常分量可能并不是只代表单一的分类特征，在 Di Tang 等人的分析中，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。后者的公式如下:

$$r = R(x) = \sigma * \mu_{t1} + (1-\sigma) * \mu_{t2} + \varepsilon \tag{2.2}$$

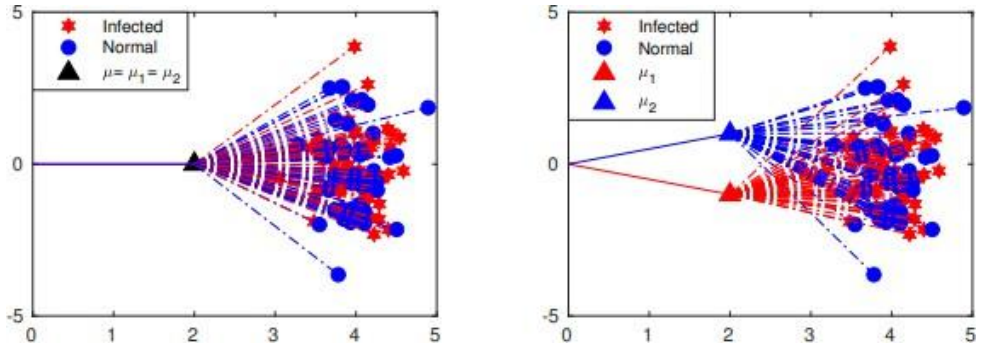

图 2.3 分别按照朴素同质假设 (左) 和双组分分解假设 (右) 对样本点分布的分析[25]

根据上述两种假设，则可以对同一个含有被感染样本的样本集合做出不同的分析。如图 2.3 所示，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，像上述基于距离的异常样本检测法一样，先标定良性样本集和其特征，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 法相对上述其他主流样本分析方法，不仅更加依赖统计特性，也具有更高的细粒度。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

上述几类人工神经网络的防御分析手段，在所依靠的理论假设上都有值得借鉴的方面。首先，SCAn 法的双组分分解法，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，也一定程度上可作借鉴，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。本文意图通过差异化训练的同质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，以及对不同模型间的预测差异的比较，来从目标测试集中提取恶意后门样本相对于良性样本的异常特征属性。

第三章　卷积神经网络的投票式模型

第一节 性能评估指标选取

投票机制的最终目的，是从污染数据集中划分出一个被推断为全部为异常污染样本的子集，所以实验的性能指标应该能够体现选定的异常污染样本子集的划分质量。

本实验选定漏报率和误报率作为异常污染样本子集的质量的评估指标。漏报率指的是全部的污染样本中未被囊括在选定子集中的比例，这个指标体现指定子集中对训练集全部污染样本的涵盖力；误报率指的是选定子集中不是污染样本的比例，这个指标则体现了投票模型对异常指标判定的有效性，也体现通过指定分析方法得到的异常指标对异常特征的反映效果。

第二节 数据集与神经网络结构的选取和调整

本实验使用 MNIST 数据集作为训练集和测试集的原型，并使用调整过的 LeNet 卷积神经网络作为目标模型。

MNIST 数据集因其规模庞大且具有随机性和代表性，在人工神经网络的各研究领域中极为常用，并且常作为分类器类型的人工神经网络的性能评估的基准测试样本集。而且因为其二维图像的性质，神经网络的后门表现也相对更加直观，因此在本实验中采用该数据集。

在程序中经过数据转换导入的 MNIST 数据集，是由张量化单通道二维黑白灰度图像和数字 GT 标签 (Ground-True Label，事实真实标签) 组成的样本-标签对的集合。其中训练集和测试集的规模分别是 60000 与 10000；单像素的灰度值使用 0 到 1 之间的浮点数表示， 0 表示全黑而 1 表示全白；数字 GT 标签为 0 到 9 之间的整数。

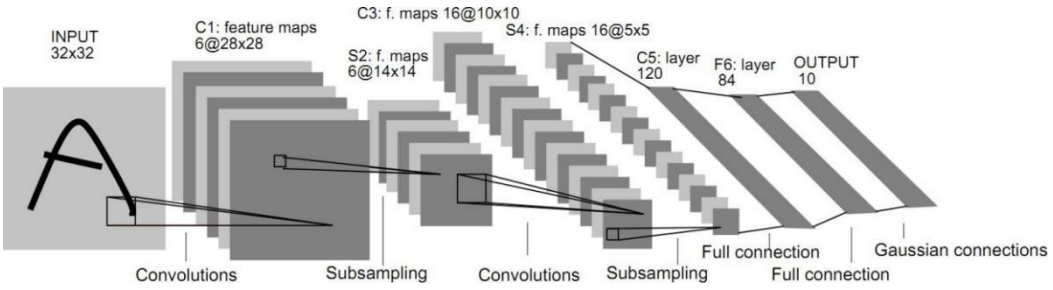

图 3.1 标准的 LeNet-5 卷积神经网络结构[19]

图3.1是典型的 LeNet-5 网络结构，但是为适应本次实验对 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和结构做出调整。调整如下:

1. 输入层: 输入层的输入张量为*batchsize*×1×28×28 的格式， *batchsize* 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，同样采用非扩张卷积法，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，输入向量和输出向量的长度分别是 320 与 50
5. 全连接层 $fc2$: 线性变换层，输入和输出向量的长度分别是 50 与 10。
6. 输出层: 使用 *logsoftmax* 函数将长度为 10 的向量函数映射为概率分布向量并取对数

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，使其并不过分依赖于局部特征，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

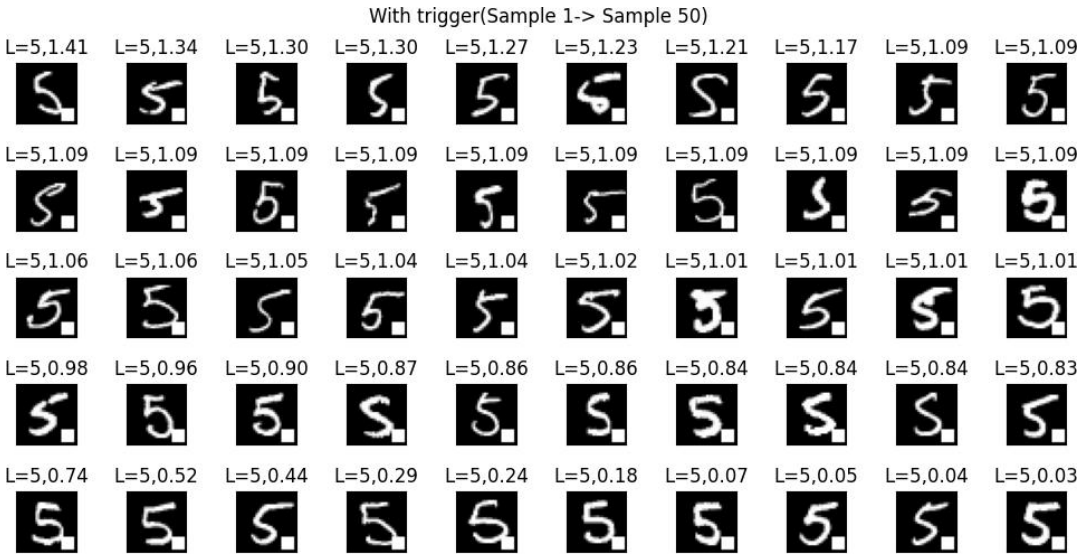

图 3.2　实验中指定 GT 标签为 5 的后门样本集示例

第三节 样本集的数据投毒预处理

、

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，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据投毒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。

对样本集的目的性数据投毒，要在网络中植入触发器后门，并同时要在测试时能使被植入后门的模型在恶意样本判别中能体现明显的输出差异，简而言之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，应当在测试集和训练集中选定无关的单一 GT 源标签类对象集合的子集作为预选的样本污染对象。

训练集中选择污染选定某单一 GT 源标签类对象集合的子集，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。

在本次实验中，实际操作上可以采取置放方块触发器的方式，对特定 GT 标签类的样本做出污染，如图 3.2 所示。若调用 python 库对 MNIST 数据集的数据集装载器处理功能，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将不合法的灰度值截断取合法的值。

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:

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

卷积神经网络对数据集一般要求使用上的随机性，需要每次都打乱样本集

并重新进行批量化分组。而且，python 相关库对 MNIST 数据集的处理依赖于官方的文件组织格式，具体如图 3.3 所示。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值即可。

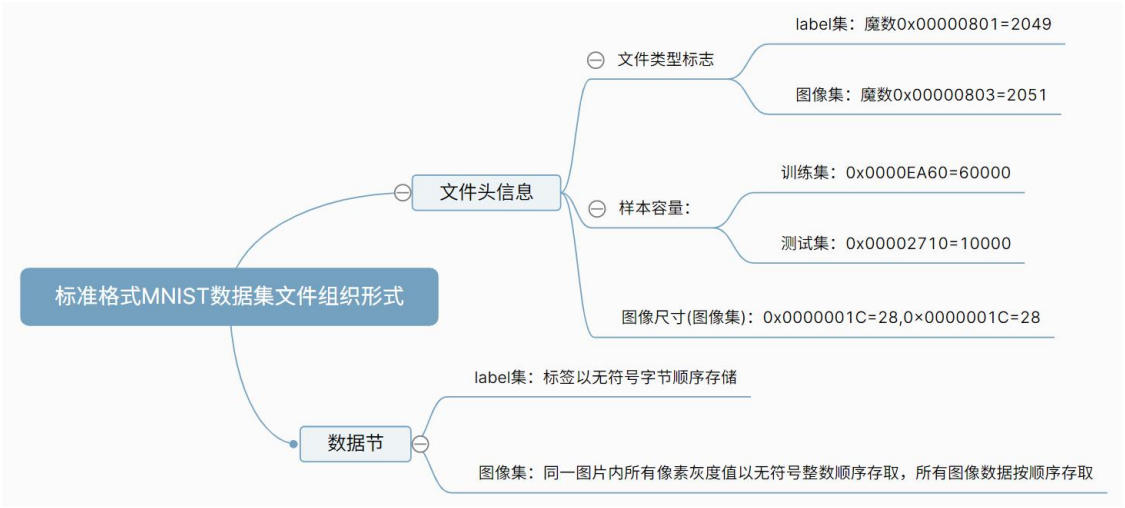

图 3.3 MNIST 标准数据集文件的数据组织规则

第四节 模型的差异化训练

模型集合的差异化训练是要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，实验中要求在预定的洁净训练模型和污染训练模型在差异训练前都具有较高的分类正确率，这可以通过给定高分类正确率的预训练模型实现。

在提高预训练模型本身的分类成功率外，体现后门植入的影响也是重要的训练目的。体现训练的差异可以从污染模型间以及洁净训练模型和污染训练模型之间分别体现。污染模型间的差异可以通过从被污染的父训练集中随机选定小规模的子训练集实现，这样可以使各模型训练集间样本重复率相对低，造成污染特征提取上的差异。而洁净模型和污染模型之间的差异则通过是否进行污染训练体现。

结合上述对污染环节的分析，在实际操作中，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，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。

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 $v_i(1 \le i \le 2n)$。v_i 的各元素是由输出层的 *softmax* 函数输出的 10 个数字类的概率值，格式如下:

$$v_i = [p_{0,i}, p_{1,i}, ..., p_{9,i}]^T \tag{3.2}$$

根据 *softmax* 函数的性质则有以下关系

$$\sum_{x=0}^{9} p_{x,i} = 1 \tag{3.3}$$

而对于样本 S 的在全模型集上得到的概率矩阵 $Matrix_S$ 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，将概率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，这里的异常指标的效用类似于 STRIP 法中的熵指标。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，需要将概率矩阵 $Matrix_S$ 按照分类切分成 10 个各模型对该样本在同一类上的概率值组成的判别矩阵子向量 v_0' 到 v_9'，如下:

$$Matrix_S = [v_0, v_1, ..., v_9]^T \tag{3.5}$$

对矩阵的子向量 v_x' 求所有分量的标准差 STD_x，函数定义为 *std*，则有:

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

然后便可得到样本 S 的异常指标 I_S 公式如下:

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

随后将测试及所有样本按照异常指标值降序排列，并取异常指标最高的 A 个样本作为指定的污染样本子集。设子集样本中最高异常指标值 I_S' 和最低值 I_S''，称后者与前者的比值为异常阈值比例 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:

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

误报率指标W公式为：

$$W = 1 - \frac{\sum_{x=1}^{N} Judge(I_{S_x})}{A} \tag{3.9}$$

若设函数T当且仅当人工识别确定样本含有触发器时值为真，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:

$$Judge(I_{S_x}) = \begin{cases} 0 \quad , \quad \dfrac{I_{S_x}}{I_S'} < p_{vaild} \\ 0 \quad , \quad T(I_{S_x}) = False \\ 1 \quad , \quad otherwise \end{cases} \tag{3.10}$$

结合上述训练和测试中求取测试集的指定污染子集的流程，可以将其总结为如下的伪代码所代表的算法，其中省略上述样本异常指标(abnormal index)的具体计算过程。

```
Start Procedure:
for every model in modelset1,modelset2: %分别指代污染模型集合和洁净模型集合
   model.parameters = pretrained_model.parameters %载入预训练模型参数
   if model in modelset1:  %仅对污染模型进行污染训练
      Subset = subset(trainset)
      model.train(Subset)  %使用污染训练集子集训练污染模型
set samplelist, resultlist empty
for every sample in testset:
   compute abnormal_index by models %计算异常指标
   append (abnormal_index,sample) to samplelist
samplelist.sort(index = abnormal_index,order = descend)  %按照异常指标降序排列
for every (abnormal_index,sample) in samplelist :
   if first:  %是该已排序集合的首样本
      top = abnormal_index %记录最高异常指标值
   if abnormal_index < top * p_valid:
      Break %当样本异常指标低于阈值时截止记录指定异常样本
   append (abnormal_index,sample) to resultlist
Return resultlist
End
```

第六节 相关参数对评估指标影响推测

由于样本投毒预处理和模型差异化训练中对参数的调整会影响投票机制的性能，本节将分析其中主要参数调整对漏报率和误报率的影响。

样本的投毒预处理中，我们主要考虑对训练集特定标签污染率和测试集污染样本数的影响。

训练集污染特定标签污染率主要是影响污染模型后门植入的效率，污染率越大后门植入越明显。因此在其它参数不变且适宜的情况下，相对高的训练集特定标签污染率会使得模型间的输出差距更大，异常样本的异常特征会更加突出，有利于异常样本子集的划分，使得漏报率和误报率更低。

而测试集污染样本数则是影响测试集中污染样本的占比。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，使得一般样本类似随机噪音的影响对异常指标区分判别的影响度相对下降，同样有利于漏报率和误报率的降低。

而在模型的差异化训练中，我们主要考虑批量规模和训练轮数的影响。

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 标签类的样本，在整个训练集中占比低且分散，而且是 MNIST 数据集本身在使用中是随机无序的，所以需要防止数据集的污染噪音化。这里所谓的噪音化，指的是由于 LeNet 卷积神经网络等常规卷积神经网络的批量化训练机制的影响，使得在大批量规模训练中，少数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，最终使得漏报率和误报率升高。为防止训练集中的后门样本噪音化，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。过高和过低的训练轮数都会使得漏报率和误报率上升。

第七节 投票机制评估指标结果分析

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

首先尝试分析训练集中指定类污染率对平均漏报率与误报率的影响。指定以下参数：洁净预训练模型数和污染模型数均取值为 3，模型子训练集规模比率 0.2，异常阈值比例 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 的测试样本集指定 m 个污染样本。通过更改参数 $P_{perturbe}$ 和 m 重复实验，得到如图3.4所示的数据。

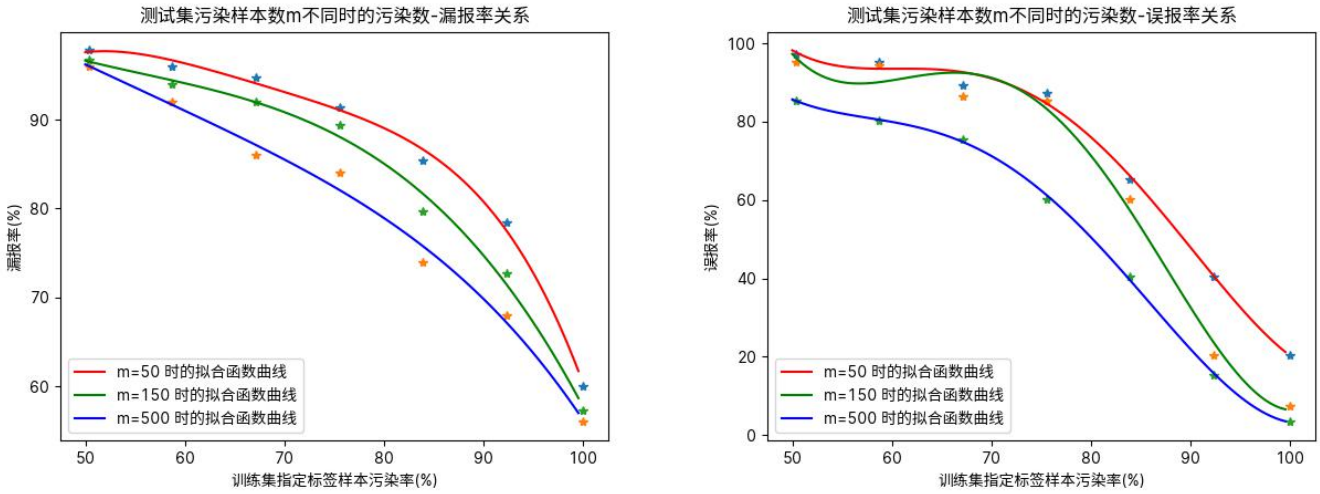

图 3.4 不同测试集污染样本数下训练集指定类污染率对应的平均漏报率和误报率

根据上面得到的数据可以得知，在各投票模型的漏报率和误报率随着训练集指定类污染率的上升而下降。而且，随着测试集污染样本数的上升，漏报率

和误报率也呈现下降的趋势。漏报率和误报率随训练集指定类污染率的变化趋势表明，污染训练的后门植入效果越明显，后门样本就越会向着高异常指标的方向集中；而随测试集污染样本数的变化趋势则说明测试集中污染样本占比的上升会冲淡正常样本对异常指标反映异常特征的能力的干扰。

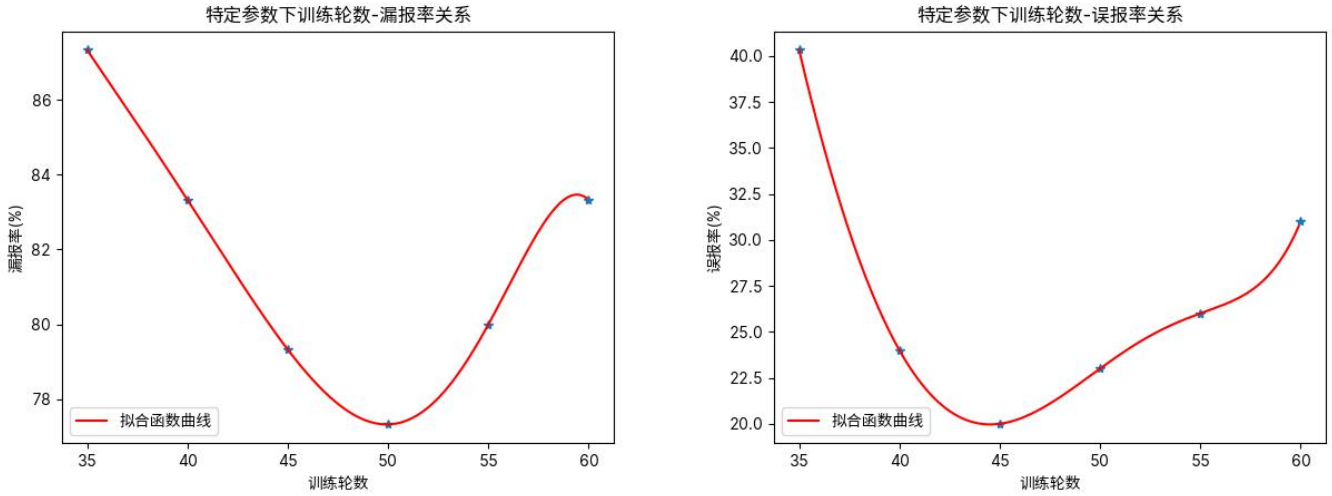

图 3.5 与图3.4其他主要参数相同时训练轮数与对应的平均漏报率和误报率

接下来尝试从实验数据分析训练轮数对平均漏报率和误报率的影响。在不更改上述主要参数的情况下，通过重复实验得到实验数据如图 3.5 所示。根据上面得到的数据可以得知，随着训练轮数的不断增长，漏报率和误报率均存在先下降后上升的趋势，这证明了在模型的污染训练中的确在训练轮数超过一定范围时存在一定程度的模型存在过拟合现象。但是在图像上训练轮数的取值域内漏报率和误报率的变化特性来看，将训练轮数控制在图像中两条函数曲线的底部拐点之间，可能在存在过拟合现象的情况下取得相对较好的训练效果。

另外，在不更改上述其他主要参数的情况下，得到批量规模与平均漏报率以及误报率关系的实验数据如图 3.6 所示。分析实验数据可以发现，模型集合的平均漏报率在批量规模低于某个阈值前随着批量规模的增加急剧上升，随后随着批量规模的增长在略低于 100% 的高平均漏报率附近振荡，而误报率的变

化趋势与之相似。这证明了模型仅在相对低批量规模下训练才能得到相对好的后门植入效果，而在高于某一阈值的高批量规模下训练集对后门样本特征的学习会接近完全失效。

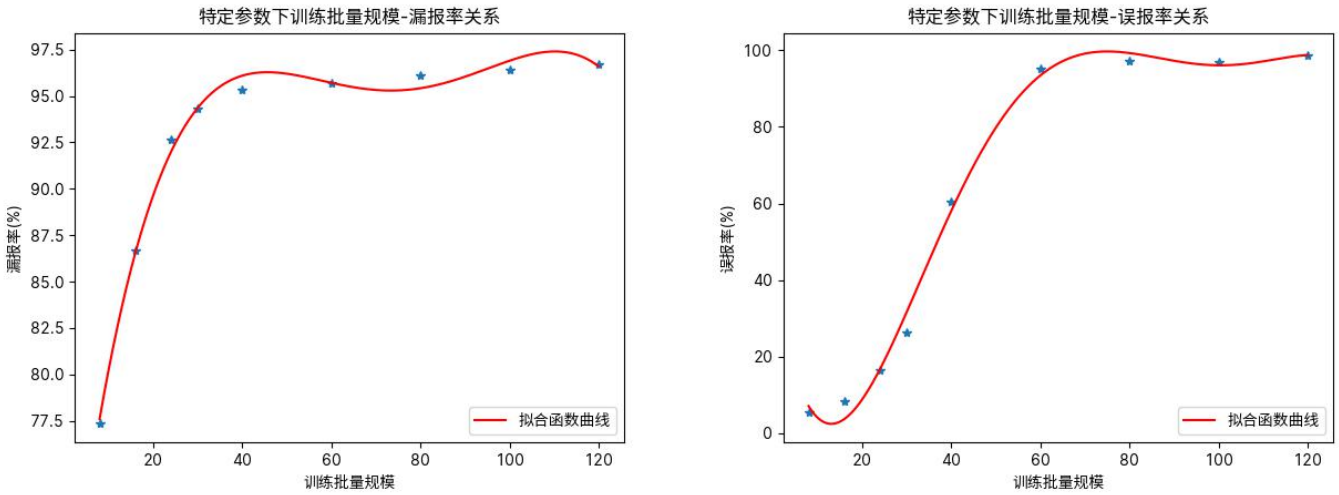

图 3.6 与图3.4其他主要参数相同时批量规模与对应的平均漏报率以及误报率

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的学习，了解其技术和概念迭代的历史脉络和应用价值。通过相关文献的查阅，了解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. 对一些与结果相关的实验参数重复调整，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. 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。

2. 本文在理念上基于对 STRIP 分析法与 SCAn 分析法在模型间输出差异上的移植，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。
3. 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。
4. 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网络架构(例如 Alex Net 等架构)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[24]、后门植入和体现输入差异有更好表现。

参考文献

[1] Culloch W, Pitts W H. A logical calculus of the ideas immanent in neural nets. Bulletin of Mathematical Biophysics, 1943, 5(4):115-133

[2] Hebb D O. The organization of behavior. 1949

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

[4] Widrow B, Hoff M E. Associative storage and retrieval of digital information in networks of adaptive “neurons”. Biological Prototypes and Synthetic Systems, 1962

[5] Minsky M L, Papert S. Perceptrons: An introduction to computational geometry. 1969

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

[7] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997, 9(8):1735-1780

[8] Lee T S, Mumford D. Hierarchical Bayesian inference in the visual cortex. Journal of the Optical Society of America, 2003, 20(7):1434-1448

[9] Serre T, Kreiman G, Kouh M, *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[博士学位论文], 2014

[11] 胡清华，张道强，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] Duda R, Hart P, Stork D. Pattern classification: Wiley-Interscience. 2000.

[13] 孙志军，薛磊，许阳明等. 计算机应用研究, 2012, 29(8):5

[14] Fukushima K. 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] Lecun Y. Generalization and network design strategies. Connectionism in Perspective, 1989

[16] 张润，王永滨. 机器学习及其算法和发展研究

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

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

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

[20] Li Y, Wu B, Jiang Y, *et al*. Backdoor learning: a survey. 2020

[21] Gao Y, Xu C, Wang D, *et al*. Strip: a defence against trojan attacks on deep neural networks. 2019

[22] Paudice A, Muñoz-González L, Gyorgy A, *et al*. Detection of adversarial training examples in poisoning attacks through anomaly detection. 2018

[23] Tang D, Wang X F, Tang H, *et al*. Demon in the variant: statistical analysis of DNNs for robust backdoor contamination detection, 2019

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