深度学习中的后门攻击综述

杜 巍, 刘功申

上海交通大学网络空间安全学院 上海 中国 200240

摘要 随着深度学习研究与应用的迅速发展, 人工智能安全问题日益突出。近年来, 深度学习模型的脆弱性和不鲁棒性被不断的揭示, 针对深度学习模型的攻击方法层出不穷, 而后门攻击就是其中一类新的攻击范式。与对抗样本和数据投毒不同, 后门攻击者在模型的训练数据中添加触发器并改变对应的标签为目标类别。深度学习模型在中毒数据集上训练后就被植入了可由触发器激活的后门, 使得模型对于正常输入仍可保持高精度的工作, 而当输入具有触发器时, 模型将按照攻击者所指定的目标类别输出。在这种新的攻击场景和设置下, 深度学习模型表现出了极大的脆弱性, 这对人工智能领域产生了极大的安全威胁, 后门攻击也成为了一个热门研究方向。因此, 为了更好的提高深度学习模型对于后门攻击的安全性, 本文针对深度学习中的后门攻击方法进行了全面的分析。首先分析了后门攻击和其他攻击范式的区别, 定义了基本的攻击方法和流程, 然后对后门攻击的敌手模型、评估指标、攻击设置等方面进行了总结。接着, 将现有的攻击方法从可见性、触发器类型、标签类型以及攻击场景等多个维度进行分类, 包含了计算机视觉和自然语言处理在内的多个领域。此外, 还总结了后门攻击研究中常用的任务、数据集与深度学习模型, 并介绍了后门攻击在数据隐私、模型保护以及模型水印等方面的有益应用, 最后对未来的关键研究方向进行了展望。

关键词 后门攻击; 人工智能安全; 深度学习

中图法分类号 TP18 **DOI 号** 10.19363/J.cnki.cn10-1380/tn.2022.05.01

A Survey of Backdoor Attack in Deep Learning

DU Wei, LIU Gongshen

School of Cyber Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China

Abstract With the rapid development of deep learning research and applications, artificial intelligence security issues become increasingly more important. In recent years, the vulnerability and non-robustness of deep learning models are continuously revealed. Numerous attack methods against deep learning models have emerged, and the backdoor attack is one of the new attack paradigms. Different from adversarial examples and data poisoning, backdoor attackers add triggers to the training data of the model and change the corresponding labels to the target class. Deep learning models are implanted with backdoors that can be activated by the triggers once they are trained on poisoned datasets. The poisoned model can still keep high precision for the normal samples while the model will output according to the target class specified by the attacker when the inputs have triggers. With this new attack scenario and setup, deep learning models show great vulnerability, which creates a great security threat to the field of artificial intelligence. Backdoor attacks have also become a popular research area. Therefore, in order to better improve the security of deep learning models for backdoor attacks, this paper presents a comprehensive analysis of the existing backdoor attack methods in deep learning. First, we analyze the differences between backdoor attacks and other attack paradigms and define the basic backdoor attack methods and processes. Then we summarize the adversary models, evaluation metrics, and attack settings for backdoor attacks. Then, the existing attack methods are classified in multiple dimensions such as visibility, trigger types, label types, and attack scenarios, encompassing various domains including computer vision and natural language processing. In addition, the tasks, datasets and deep learning models commonly used in backdoor attack research are summarized, and useful applications of backdoor attacks in data privacy, model protection, and model watermarking are presented. Finally, the key research directions in the future are prospected.

Key words backdoor attack; artificial intelligence security; deep learning

1 引言

近年来, 机器学习和深度学习模型在实际生活中的应用越来越多, 例如人脸识别[1]、自动驾驶[2]、机器翻译[3]、语音处理[4] 等。这些模型大都需要庞大的数据量进行训练才能够达到远超传统方法的效

通讯作者: 刘功申, 博士, 教授, Email:lgshen@sjtu.edu.cn。

本课题得到国家自然科学基金项目 (No. 61772337) 资助。

收稿日期: 2021-03-09; 修改日期: 2021-04-30; 定稿日期: 2022-03-21

果。与之相应的另一个问题便是计算资源, 机器学习尤其是深度学习的模型架构往往十分庞大, 具有大量的参数, 例如自然语言处理中 GPT-3 模型的参数量达到了 1750 亿[5], 这对深度学习模型的训练提出了非常高的要求。许多研究者和使用者不具有足够的数据和计算资源, 通常会选择与第三方合作, 通过使用第三方提供的云计算平台来完成深度学习模型的训练, 由此也诞生了许多针对机器学习与深度学习模型训练的云平台服务 MLaaS (Machine Learning as a Service)[6], 或是下载和使用一些开源或第三方提供的数据和预训练模型。此外, 一些用户不具有深度学习的相关知识, 可能会直接部署和使用第三方训练好的模型。

这种合作训练机器学习模型的方式, 使得模型的训练过程完全暴露在第三方平台中, 用户失去了对训练过程的完全控制, 假如第三方平台存在恶意目的, 那么恶意攻击者可以轻松的对模型进行修改和破坏, 造成严重的后果。例如, 文献[7]中展示了一个遭受攻击的交通标志识别模型, 如图 1 所示, 攻击者在“停止”交通标志上添加了一个不显眼的便利贴, 使“停止”标志被模型识别为“限速”标志。由此可见, 人工智能安全问题越来越接近现实生活, 针对深度学习模型的攻击越来越具有威胁, 这引起了广大研究者的关注。

图 1 深度学习模型识别停止标志

Figure 1 Deep learning model recognizes STOP

后门攻击是针对上述场景的新型攻击范式, 近年来的相关研究激增, 但针对该方向的总结性研究较少, 因此本文调研了该方向的相关进展, 希望以一个全面而综合的视角, 对后门攻击领域进行分析与总结。

本文剩余部分的组织结构如下: 第 2 章对比了包括后门攻击在内的深度学习模型的攻击范式; 第 3 章阐述了后门攻击的定义, 并对相关的基本知识进行了总结归纳; 第 4 章根据不同后门攻击方法的特点进行了分析和归类; 第 5 章总结了后门攻击中常用的数据集与深度学习模型; 第 6 章从多个角度对未来的关键研究方向进行了展望; 第 7 章总结全文并给出结论。

2 深度学习模型攻击范式

现阶段, 针对前述场景中深度学习模型的攻击手段主要有对抗样本攻击[8-11]、数据投毒攻击[12-15] 以及后门攻击[7], 三者存在一定的区别, 如表 1 所示。

对抗样本攻击主要存在于模型的推理阶段, 针对一个已经训练好的模型, 希望构造一个能够欺骗模型的样本, 而不会修改和破坏已有的模型。后门攻击和数据投毒攻击则主要存在于模型的训练阶段, 都是通过对训练数据进行修改也即投毒, 对模型产生影响和破坏。

后门攻击与数据投毒攻击的不同之处在于攻击目的, 数据投毒的主要目的是使模型的泛化性能变差, 也即在测试集上的效果变差, 模型不能进行有效的学习, 甚至无法收敛。而后门攻击的目的则是使模型学习到攻击者指定的内容, 其对正常样本仍旧具有良好的测试效果, 但对于中毒样本则会输出攻击者预先设定的标签。

表 1 不同攻击范式对比

Table 1 Comparison of different attack paradigms

攻击类型	攻击阶段	对模型的影响
对抗样本	推理阶段	欺骗模型
数据投毒	训练阶段	破坏模型
后门攻击	训练阶段	诱导模型

相比于其他攻击, 后门攻击更具威胁, 其原因主要有以下几点:

复杂性: 对抗样本攻击主要研究模型推理阶段对于对抗样本的脆弱性, 而与推理阶段相比, 模型在训练阶段涉及更多的步骤, 包括数据采集、数据预处理、模型构建、模型训练、模型保存、模型部署等等[16]。更多的步骤意味着攻击者有更多的机会, 模型的安全威胁也更多。

隐蔽性: 后门攻击对于正常样本来说没有异常, 只有当样本具有后门触发器时才会发生异常, 因此用户在使用时难以察觉, 此外, 后门攻击注入的中毒样本通常非常之少, 仅需 0.5%左右[17]。

实际性: 数据投毒攻击希望模型在测试集上效果变差, 而在实际中, 对于正常测试集效果较差的模型通常不会投入使用。相反, 后门攻击保证模型在

正常测试集上仍具有良好效果, 因此经过后门攻击的模型很大概率会部署并投入使用。

3 后门攻击定义

3.1 术语和标记

本节针对深度学习后门攻击中的术语进行定义和解释, 并给出在后文中使用的对应标记。

1) 正常样本 x_i: 未经后门攻击的原始数据;

2) 中毒样本 x_b: 通过后门攻击手段得到的样本数据, 可以通过训练将后门埋藏在模型中, 通常是对正常样本进行修改得到;

3) 源标签 y_i: 中毒样本对应正常样本的标签;

4) 目标标签 t: 攻击者所指定的用于埋藏后门的类别标签, 通常是使模型误分类的类别;

5) 正常数据集 D: 不含中毒样本的原始数据集;

6) 中毒数据集 D_b: 注入了中毒样本的数据集;

7) 正常模型 M: 通过正常数据集训练的模型;

8) 中毒模型 M_b: 通过中毒样本训练而被埋藏了后门的模型;

9) 触发器/后门模式Δ: 后门攻击中用来生成中毒样本和激活模型后门的一种模式。

3.2 攻击定义

后门攻击方法 $f(\cdot)$基于触发器或后门模式Δ对正常样本 x_i进行处理, 得到中毒样本 x_b, 即 $x_b=f(x_i, \Delta)$, 并为该中毒样本指定目标标签为 y_t, 然后将多个中毒数据对(x_b, y_t) 和正常数据(x_i, y_i)一起组成新的训练数据集, 用来训练神经网络模型, 得到埋藏了后门的模型 M_b。当使用该模型对正常样本 x_i^{test} 进行预测时, 模型仍然可以得到正确的预测结果 $M_b\left(x_i^{test}\right)=y_i$, 而当使用该模型对带有触发器$\Delta$的中毒样本 x_b^{test} 进行预测时, 模型会按照攻击者所指定的目标类别标签输出, 即 $M_b\left(x_b^{test}\right)=y_t$。

3.3 敌手模型

本节从敌手知识和敌手能力两个方面来描述后门攻击的敌手模型。

敌手知识: 从理论上来说, 后门攻击者可以获得的知识包括训练模型使用的正常数据以及模型内部的架构和参数。但在实际情况中, 攻击者通常可以通过外包或者第三方来收集相关的训练数据, 但很难直接访问到模型内部, 因此大多数后门攻击方法都是基于正常数据及其标签而展开的[7, 17-18]。

敌手能力: 根据敌手知识, 攻击者可以修改模型的内部结构或参数进行后门攻击[19-21], 或是根据模型的内部参数来构造生成中毒样本[18, 22], 但在实际情况中更多的是通过修改正常数据及其标签来达到攻击目的[7, 17-18], 还有不少攻击者可以在不修改中毒数据的标签的情况下进行后门攻击[23-25]。此外, 攻击者通常可以控制模型的训练过程, 通过从头训练或重训练来注入后门。

根据后门攻击者获得的知识与可以使用的能力, 可以将攻击方法划分为黑盒模型和白盒模型。白盒模型的方法更为常见, 其允许攻击者可以访问或获取模型的训练数据。黑盒模型则要求攻击者在无法获取到模型的训练数据的情况下进行攻击, 这种情况更接近于真实情况。通常, 黑盒模型的后门攻击根据某些方法来生成一些训练样本, 然后再使用白盒模型的方法进行攻击[18]。

3.4 评估指标

对于深度学习模型中的后门攻击来说, 主要通过以下三个指标进行评估[26]。

攻击成功率 (Attack Success Rate, ASR): 指成功使模型误分类为目标类别的中毒样本所占的比例。

准确率下降 (Accuracy Decline, AD) : 指模型在后门攻击前后, 对于正常样本预测准确率的下降值。

攻击隐匿性 (Attack Stealthiness, AS): 指后门攻击方法躲避人类视觉检查以及一些检测方法的能力。

ASR 和 AD 针对模型的表现而言, 通常来说, 模型经过后门攻击后, 对于正常样本预测的准确率会下降, 而准确率下降越少, 越不易引起使用者或防御者的察觉, 模型也会更可能部署使用。因此攻击者希望尽可能减小对模型正常性能的损害, 使 ASR 尽量高而 AD 尽量低。

AS 则从攻击方法本身的隐蔽性或不可见性出发, 对后门攻击方法的设计提出要求。为了躲避人类视觉检查或一些检测方法, 通常需要对触发器的形状、大小、透明度以及投毒率等进行限制。为了量化体现 AS, 可以定义例如数值变化率、结构相似性[27-28]等指标进行评估。

4 深度学习后门攻击

本章总结了目前深度学习后门攻击的相关研究进展, 并针对各后门攻击方法的特点对其进行了分析与归类。首先介绍了后门攻击存在的各种攻击模式, 之后对本文所归类的计算机视觉领域的 12 类后门攻击进行了阐述并总结了相关文献, 同时对其他相关领域的后门攻击进行了总结。表 2 是对其中关

键研究的归纳, 总结了其中的相关实验。

4.1 攻击设置

后门攻击可以从不同方面的进行设置, 而不同的设置各有侧重, 本文总结如下。

4.1.1 触发器

1) 触发器属性: 主要包括触发器的形状、大小、位置以及透明度。触发器的形状与大小会对原数据的特征产生影响, 对于原特征覆盖面积越大的触发器, 会使模型更倾向于学习触发器特征而忽略原数据特征, 但同时也使其视觉隐匿性变差; 触发器的位置信息也会影响后门植入的过程, 可以控制后门只可在固定位置触发还是任意位置皆可触发; 触发器的透明度用于衡量触发器和原始数据之间的混合程度, 通常透明度越高其视觉隐匿性越高, 但同时模型也越难以学习到触发器特征。

图 2 触发器属性

Figure 2 Trigger properties

2) 触发器类型: 主要包括确定图案、动态图案以及良性特征的 3 种类型。确定图案指使用攻击者设计的固定图案作为触发器; 动态图案指具有输入感知功能的攻击方式, 可以根据不同输入产生不同图案的触发器[29-30]; 良性特征则与上述两种不同, 其不植入额外的特征, 而是使用原数据本身具有的良性特征作为触发器[28,31-32],例如人脸面部特征等。

3) 攻击类型: 根据触发器的使用方式可以分为 3 种类型。单对单攻击指单个触发器激活单类目标后门; 多对单攻击使用多个触发器, 当多个触发器同时触发时才激活某单类目标的后门; 单对多攻击则使用同一触发器, 根据不同的触发强度来激活不同目标的后门。

4.1.2 目标类别

1) 单对单攻击: 仅使某一类别的数据, 在添加触发器后被分类为目标类别, 其他类别添加触发器后仍正常分类。单对单攻击希望模型学习某一类数据和触发器的特征组合与目标标签之间的联系。

2) 多对单攻击: 使所有或多个类别的数据, 在添加触发器后被分类为目标类别。多对单攻击则希望模型学习触发器本身的特征, 从而使模型对所有带有触发器的数据都按照预定的标签输出。

4.1.3 训练方式

1) 从头训练: 使用中毒数据集对模型从零开始进行训练, 通常耗时较长, 但效果通常较好。

2) 微调: 使用中毒数据集对已在正常数据集上训练好的模型进行重训练, 耗时较短, 但有时效果一般。

此外, 还有一些方法不使用训练的方式注入后门, 而是直接篡改模型参数,可以达到与训练注入后门同样的效果[19-21]。

4.2 计算机视觉领域的后门攻击

后门攻击最早针对图像数据提出, 因此大多数后门攻击研究都针对计算机视觉领域展开, 本节对近年来的相关研究进行了分类总结。如图 3 所示, 分别从可见性、触发器类别、标签类别、攻击形式、场景设置进行具体分类。

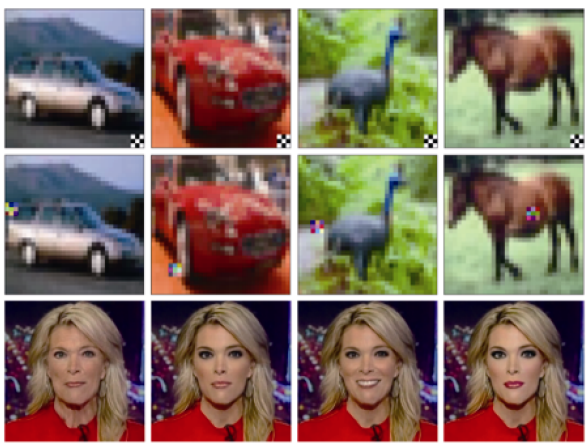


图 3 触发器类型

Figure 3 Trigger type

4.2.1 简单后门攻击

Gu 等人[7]首先提出了深度学习模型后门攻击的概念, 是后门攻击领域的开山之作, 其描述了后门攻击的基本步骤, 首先在正常数据上添加触发器作为中毒数据, 然后为中毒数据打上攻击者指定的目标标签, 最后将这些中毒数据与正常数据一起训练, 训练的目的为, 对于带有后门模式的样本, 模型学

习到的是后门模式的特征, 从而可以将任意带有后门模式的数据分类为目标标签, 对于正常数据, 模型仍然学习数据本身的特征, 正常分类。类似的, Guo 等人[33]提出了一种针对人脸匹配系统的简单后门攻击, 称为通用身份攻击, 可以使用某个特定的人脸冒充任意的合法人脸类别。作者简单的对正常数据进行修改, 替换两张人脸数据中的一张人脸为特定人脸作为激活后门的模式, 并改变标签为合法, 然后使用中毒数据集对网络进行训练。

4.2.2 不可见后门攻击

简单后门攻击使用异常明显的后门模式显然不符合实际要求, 人类可以轻易的找出中毒数据。因此不少学者开始研究不可见后门攻击。不可见后门攻击指后门模式不可见, 使人类难以区分中毒数据与正常数据。显然, 当触发器与图片之间难以区分时, 模型将更难学习到触发器的特征, 因此如何设计一个能够逃过人类检查但又能使模型能够充分学习的“隐形”触发器成为一个热门研究方向。

Chen 等人[17]首先探索了这一问题, 不同于简单后门攻击中的直接替换[7], 其将后门模式以一定程度叠加在原始图片的像素值上, 并采取了两种不同的后门注入模式。Input-instance-key 模式在图像的数字空间内使用一定幅度的随机噪声作为触发器, 尽管由于随机性使得在训练和测试时使用的噪声不同, 但仍取得了较高的攻击成功率; Pattern-key 模式基于某种给定样式的触发器以一定比例与原图片混合。此外, 还使用紫色太阳镜和黑框眼镜等人脸配件作为触发器, 模拟了真实世界的攻击。Turner 等人[23]在针对图像的数据增强中也做了类似的透明处理。

Liao 等人[34]提出使用不可见的对抗生成扰动作为触发器进行后门攻击, 并采取了两种生成扰动后门模式的方法。具体的, 模式静态扰动指构造一种简单的小幅度重复图案作为触发器, 目标自适应扰动是指使用一种通用对抗性扰动[35]作为后门触发器, 该扰动使输入靠近目标标签的决策边界, 并且约束其大小以保证不可见性。

Li 等人[27]针对不可见后门攻击提出了两种方法, 第一种基于图像隐写技术, 使用经典的 LSB 算法将触发器嵌入到比特位空间中, 第二种则类似于文献[34], 将 L_p 正则化约束得到的扰动增量作为触发器。

Nguyen 等人[36]则认为人类可以识别出图片中不一致的部分, 在图片上添加扰动噪声、条纹或反射等方式难以躲过人工检查, 而人类不擅长识别较小的几何变换, 因此提出以保留图像内容的微小扭曲形变作为触发器, 使中毒图像更真实与自然, 从而更容易躲过人工检查。该文构建了 WaNet 来生成具有扭曲形变的中毒图像, 首先对随机噪声进行上采样与裁剪生成用于图像形变的二维扭曲场, 然后将其应用到正常图像上, 产生人类难以察觉的微小形变。

最近, Sarkar 等人[31]使用面部属性或特定表情作为触发器成功实施了针对人脸识别系统的不可见后门攻击。该文对于触发器进行了两种尝试, 一是使用 FaceApp 中的一些滤镜对面部特征进行修改作为触发器, 即人工改变的面部特征, 二是尝试使用原有的面部特征或面部表情作为触发器, 即自然存在的面部特征, 例如微笑、眉毛挑起、眯起眼睛、嘴唇微张等。而 Xue 等人[28]认为使用如太阳镜等配件[17]或一些特定表情[31]作为人脸识别系统的触发器不适用于真实场景, 因此针对人脸识别系统提出了两种新的不可见攻击 BHF2 和 BHF2N, 该方法将触发器隐藏到人脸的胡须和眉毛中, 攻击者只需通过简单的化妆即可触发后门攻击。

4.2.3 干净标签攻击

不可见后门攻击虽然使中毒图片类似于正常图片, 但其标签仍然不同于正常版本。通过检查训练样本图像与标签的关系, 仍可以检测到这种不可见的攻击。由此衍生出了干净标签攻击这一研究方向, 其要求中毒数据标签与真实标签保持一致的前提下, 让模型在学习到目标类原始数据特征的同时, 也可以学习到后门模式的特征, 使后门特征成为模型输出目标类的充分不必要条件。

Barni 等人[24]针对干净标签攻击进行了简单的探索, 给出了干净标签攻击的基本设置, 即不更改中毒样本的标签, 仅使目标类的部分样本中毒。并且通过实验说明, 相比于简单后门攻击与不可见后门攻击通常只需要注入 1%~5%的中毒样本, 干净标签攻击需要增加中毒样本比例至 20%以上。

Turner 等人[23]则提出了一种更加有效的攻击方法, 注意到当中毒数据的标签与真实标签保持一致时, 模型自然的会更倾向于学习图像本身的特征, 而忽略后门模式的特征, 因此该文章考虑, 在不明显改变图像的前提下, 模糊图像本身的特征, 使模型更难通过学习图像本身特征进行分类, 迫使模型学习后门模式的特征进行分类。具体的, 该文通过两种方式来模糊图像特征, 一种是基于 Gan 在原图片中插入其余类别的图片信息来混淆特征, 另一种是利用对抗扰动的思想, 使用基于 L_p 约束的 PGD 方法在图像中引入扰动。

Saha 等人[25]针对不可见和干净标签提出了新的攻击方式, 对于某个中毒样本, 考虑使其在像素空

间中尽可能接近目标类别的样本, 而在特征空间尽可能接近添加了触发器的原样本, 这样就可以在躲过人类的检查同时也使模型学习到触发器特征。Ning 等人[37]则考虑对触发器进行表示, 其使用自编码器将原始触发器转换为一种人类不可见的噪声图像, 该噪声图像具有与原始触发器相同的特征表示, 从而在注入后门时可以达到相同的效果。

现有的后门攻击算法大多是针对图像载体的, 由于视频具有更高的维度与更稀疏的数据场景, 将针对图像的攻击方法用到视频上效果会显著下降, 因此 Zhao 等人[38]探索了针对视频任务模型的后门攻击方法, 提出了一种通用对抗扰动触发器来为视频任务模型注入后门。为了使该方法在干净标签的情况下成功, 该文沿用了文献[23]中的方法, 对用于训练的目标标签图像也加入对抗性扰动, 模糊图像特征。该文还将该方法用来改进对于图片的后门攻击, 成功提高了高分辨稀疏图像的后门攻击成功率。

最近, Quiring 等人[39]受图像缩放攻击[40]启发, 将该方法应用到后门攻击中。图像数据在深度学习任务中通常需要进行 resize 等缩放操作来调整不同图像到同一大小, 而文献[39]中发现通过简单操作可以使图像缩放后表现为其他不相关的图像。因此, 该文利用这一点将后门模式隐藏在图像中, 将添加了触发器的图像通过缩放伪装成目标类的图像, 可以在干净标签的条件下向数据集中注入中毒数据, 具体流程如图 3 所示。

4.2.4 特定标签攻击

特定标签攻击指仅当触发器添加到某一种标签的数据上时可以激活模型后门, 触发器添加到其他标签的数据时无法激活后门。从躲避检查的角度看, 大多数防御手段都建立在触发器与样本不相关的假设上, 而特定标签攻击将触发器与目标标签类样本相联系, 因此这种后门攻击可以躲避大多数检测不同样本的相同中毒行为的防御手段。从模型角度看, 特定标签攻击是要求模型学习触发器与特定类样本的特征组合与目标标签的联系, 而不仅仅是触发器特征与目标标签的联系。

Li 等人[41]首先探索了特定标签攻击, 该文认为当前后门防御手段都是在触发器与样本无关联的假设下进行的, 例如 Neural Cleanse[42]、Fine-pruning[43]和 STRIP[40]。如果使触发器与特定样本关联将使攻击更加隐蔽。受深度学习网络图像隐写[44-45]的启发, 该文在特定类攻击的场景下, 使用编码器-解码器网络将攻击者指定的字符串编码到良性图像中作为后门模式, 构造中毒数据并进行训练将后门模式编码到模型中。编码器用来构建中毒图像, 训练目标为最小化中毒图像与正常图像之间的差异。解码器用来解码中毒图像中的触发器, 训练目标为最小化编码的重建损失。

4.2.5 反向工程

基于反向工程的后门攻击主要在于研究触发器与模型神经元权重之间的关系。可以固定神经元权重调整触发器, 也可以固定触发器调整神经元权重。

Liu 等人[18]首先探索了该方向, 根据模型的反向工程, 获取优化的触发器, 并且在无法访问真实数据集的情况下构造攻击样本, 实现后门攻击。具体的, 该文首先调整触发器像素使模型中间层的某神经元激活, 目的在于获取一个能够与所选神经元建立强连接的触发器。然后从该任务相关的公开数据集中寻找一些数据, 将数据分别对应各个最后一层的输出神经元, 调整输入图像的像素使该输出神经元激活, 目的在于获取一些能够与目标输出神经元建立强连接的训练数据, 是对真实数据集的模拟。最后使用生成的训练数据来重训练模型, 目的在于建立中间神经元与目标输出神经元的强连接, 也即触发器与目标标签之间的强连接。Yao 等人[46]也使用了类似的思想, 将其在应用到迁移学习中, 通过反向工程生成与迁移学习中冻结的中间层相关联的触发器, 从而使后门在迁移时得以保留。

最近, Cheng 等人[22]提出了一种深度特征空间后门攻击, 与常见的固定像素触发器不同, 该文使用人类难以解释的深度特征作为触发器, 对于不同的输入具有不同的表现, 通过控制解毒的方式, 抑制模型提取简单的特征而得到深度特征。具体的, 首先使用如图 4 所示的 CycleGAN 对正常图片进行风格转移作为初始触发器, 然后在一个已经达到较高攻击成功率的中毒模型进行反向工程, 通过改变触发器输入控制中毒神经元 输出较高激活值, 使模型可以从简单触发器特征中解毒, 然后再使用解毒得到的触发器对模型进行重训练, 中毒和解毒的过程反复进行, 最终得到代表了深度特征的触发器, 使得模型不依赖于简单特征进行触发。

4.2.6 多后门攻击

多后门攻击根据攻击模式一节中的触发器攻击类别角度, 分为单对多后门攻击和多对单后门攻击。

与以往的单个触发器或后门针对某一类目标不同, Xue 等人[47]提出了两种新的攻击方式, 单对多攻击与多对单攻击。具体的, 单对多攻击通过调整触发器的像素值大小来对不同的目标类别注入后门, 考虑到强度高的触发器模型更容易学习, 因此对于强

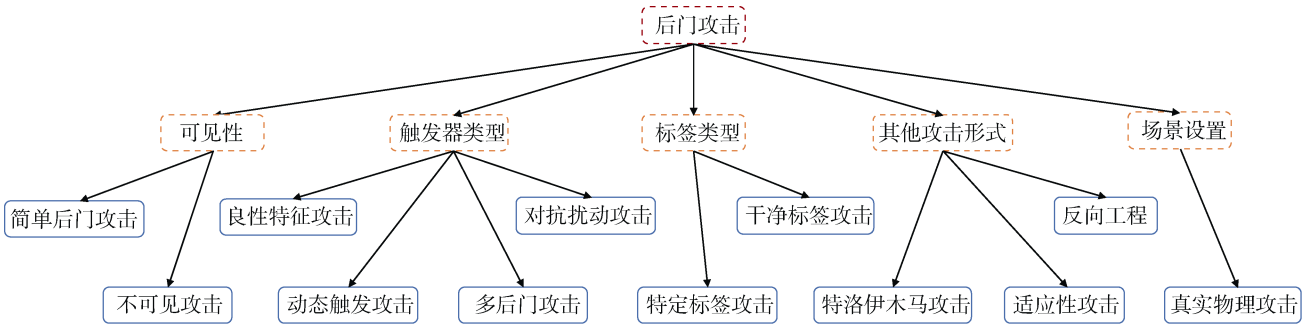


图 4　后门攻击分类

Figure 4　Backdoor attack classification

图 5　不可见后门攻击对比

Figure 5　Invisible backdoor attack comparison

度较弱的触发器应当注入更多的中毒样本来保证攻击成功率。多对单攻击则要限制包含单个触发器样本的数量，从而使模型无法学习到单独的触发器，当仅有一个触发器进行触发时模型会以较低的置信度输出目标类别，而更倾向于按照真实类别输出，因此一个触发器无法激活后门，而当多个触发器同时触发时，产生“累加效应”，模型就会以较高的置信度输出目标类别。相比于单对单的形式，这两种攻击方式更为隐蔽和灵活，防御者通常只能检测到其中一个目标或触发器，很难将其完全移除。

4.2.7　良性特征攻击

绝大多数后门攻击都是在正常样本上加入与样本特征不同的触发器特征，让模型学习到触发器特征与目标标签的强联系，因此不少后门防御手段基于这一点进行检测和消除，取得了很好的效果。良性特征攻击则针对这一问题进行改进，考虑使用正常样本中存在的特征作为后门触发器，而不添加新的特征，从而使后门攻击躲过那些遵循触发器特征与良性特征之间差异进行防御的方法。

Lin 等人[32]使用多个正常标签类别的特征进行组合作为后门模式，例如人脸识别中任意两张人脸组合即可误导模型输出目标标签，这种攻击方法可以使大多数防御手段失效，如 Neural Cleanse[42]和 ABS[48]。该文提出，尽管将作为触发器的两张人脸添加进任意正常样本中都可以导致模型误分类，但在训练时无需使用触发器和正常样本叠加形式的中毒样本，而是直接使用触发器作为中毒样本进行训练，这样可以减少其他良性特征的影响。

Bagdasaryan 等人[49]将后门攻击转换为多任务学习的过程，模型在学习原任务的同时学习后门任务，并且可以无需修改原数据集，即进行良性特征攻击，用同一数据集进行两个任务，例如原任务为计算图片中的人脸个数，而后门任务为识别某个特定的人脸。Sarkar 等人[31]尝试使用原有的面部特征或面部表情作为触发器，例如微笑、眉毛挑起、眯起眼睛、嘴唇微张等，对人脸识别系统进行后门攻击。Xue 等人[28]则将触发器隐藏到人脸的胡须和眉毛中，可以作为良性特征触发器进行后门攻击。

4.2.8　对抗扰动攻击

对抗扰动攻击指借助对抗样本中的思想，将扰动作为触发器来注入后门，通常该扰动满足正则化约束使扰动不可见。

Liao 等人[34]首先使用对抗性扰动作为触发器进行后门攻击，并且约束扰动大小以保证不可见性。随后，Li 等人[27]研究了基于 L_p 正则化约束对抗性扰动作为触发器的攻击方法，图 6 给出了不同扰动下的中毒图片。类似于文献[34]中的对抗性扰动方法，Garg 等人[50]也提出使用对抗性扰动作为触发器实施后门攻击，但不同于文献[34]中针对输入空间的微小扰动，该文关注对抗性扰动对于模型神经元权重的

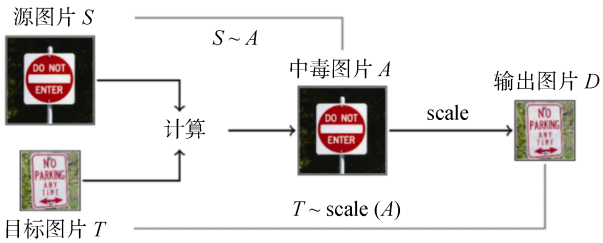


图 6　图像缩放攻击

Figure 6　Image-scale attack

影响, 考虑对模型权重进行对抗性扰动来注入后门。通过对模型权重空间使用 L_∞ 约束的梯度投影下降来添加对抗性的微小扰动, 目标是使模型仍对正常数据保持良好的准确度, 而对添加了触发器的中毒样本输出目标标签。

此外, Turner 等人[23]使用对抗性扰动来模糊原图像特征, 从而迫使模型能够在干净标签下学习触发器特征。最近, Zhao 等人[38]提出了一种通用对抗扰动触发器来为视频任务模型注入后门, 并沿用了文献[23]中的方法, 对用于训练的目标标签图像也加入对抗性扰动, 加强模型对触发器特征的学习效果。

4.2.9 适应性攻击

与对抗学习类似, 后门学习中也有研究者开始研究自适应的后门攻击, 通过归纳后门防御方法的基本原理, 然后将其纳入到模型训练的损失函数中, 训练出对于后门防御具有鲁棒性的中毒模型, 从而使模型自适应的躲避后门防御算法。

Tan 等人[51]观察到大多数后门防御算法都根据中毒样本和正常样本的特征表示之间所存在的差异进行检测和防御, 因此提出一种对抗式的后门嵌入方法, 通过对抗正则化来最大化中毒样本和正常样本之间潜在的不可区分性, 使植入了后门的模型可以广泛而有效的对抗一般的后门防御算法。基于上述思想, 该方法的在训练模型时的目标包括模型正常的分类损失以及特征表示差异损失。特征差异损失可以针对攻击者预期的某种防御方式来特别设置, 也可以设置针对各种防御方法的一般损失项。该文针对 Neural Cleanse[42]中根据平均激活数来修剪神经元的方法, 将对于正常样本与中毒样本的神经元的平均激活数的差异作为损失项, 从而可以避过剪枝类的防御方法的。

Costales 等人[52]针对 STRIP[53]防御方法进行了适应性处理。STRIP 方法扰动各类输入计算 softmax 层的平均熵, 而扰动后的中毒样本的熵较小, 可以通过设定熵阈值可以将中毒样本过滤, 因此该文作者将正常样本与中毒样本间熵分布的差距作为正则项纳入损失函数, 在重训练的过程中保持熵分布, 从而逃避 STRIP 的检测。Bagdasaryan 等人[49]也尝试了适应性攻击, 针对 Neural Cleanse[42]、SentiNet[54]以及 Gradient-Shaping[55]这三种防御方法进行了测试。

最近, Ali 等人[56]为了躲避一些进行特征检查的后门防御方法, 提出了两种低置信度后门攻击, 分别为ε-攻击和ε^2-攻击。ε-攻击将中毒样本的标签改为具有较低置信度的概率分布, 使中毒样本有更小的梯度和与干净样本更相似的特征表示。ε^2-攻击使用两个触发器分别针对两个不同的标签, 同样使用低置信度的概率分布标签, 并且在两个触发器同时出现时表现为其中一种触发器的效果, 从而将另一种触发器在特征空间中隐藏, 更好的躲过后门防御手段的检测。

4.2.10 动态触发攻击

后门攻击通常使用具有固定模式和固定位置静态触发器, 这种触发器一般难以躲过许多已有的后门防御手段, 且在真实物理世界也会受到诸多因素的干扰, 例如摄像机拍摄时的角度、光线等, 从而降低了后门攻击的实用性和威胁性。动态触发攻击则针对这一问题, 提出了触发器动态化的概念。

Nguyen 等人[29]首先针对动态触发攻击进行研究, 提出了一种新颖的输入感知后门攻击方法, 可以根据不同输入来产生不同的触发器模式, 使触发器动态化, 并且不同输入产生的触发器仅在该输入有效, 对于其他输入无效。这区别于绝大多数使用统一触发器进行触发的后门攻击, 动态且唯一的触发器使得该方法更为隐蔽。该方法使用自编码器训练针对输入的触发器生成器, 通过多样性损失保证了生成触发器的差异性, 通过交叉触发损失保证了触发器的唯一触发性。

Salem 等人[30]针对固定模式和位置的静态触发器, 研究了后门攻击中动态触发器的可用性, 通过生成对抗模型来生成动态触发器, 生成的触发器具有随机的位置和模式, 可以攻击具有不同触发模式的相同标签, 这些触发模式具有相同的潜在特征表示。

4.2.11 真实物理攻击

从使用场景上, 后门攻击可以分为来自数字空间的攻击和来自物理世界的攻击。数字攻击指触发器是来自样本输入空间中的某些扰动, 通常针对图像中的某些像素。物理攻击则指触发器是物理世界中真实存在的某些物体, 例如人脸上的眼镜或交通标识牌上的便利贴等。显然, 来自物理世界的后门攻击更具威胁, 攻击者可以轻而易举的激活后门, 同时也对后门攻击方法提出了更高的要求, 需要对物理世界的一些光线、视角、噪声等干扰具有泛化能力。

Chen 等人[17]最先探索了物理场景下的后门攻击, 使用黑框眼镜和紫色太阳镜作为人脸图片上的触发器, 将触发器以一定程度叠加在原始图片的像素值上, 取得了不错的效果。Bagdasaryan 等人[49]将后门攻击转换为多任务学习, 用同一数据集进行原任务

与后门任务, 通过代码中毒的方式分别使用了如图 7 所示的后门模式进行了尝试。Liu 等人[57]受文献[58]中基于物理反射现象的对抗样本的启发, 提出了一种反射后门攻击 Refool。利用物理反射的数学模型得到干净图像的反射图像, 将反射图像作为触发器, 使图像看起来更接近于真实物理世界的情况, 并且在干净标签的情况下达到了较好的效果。

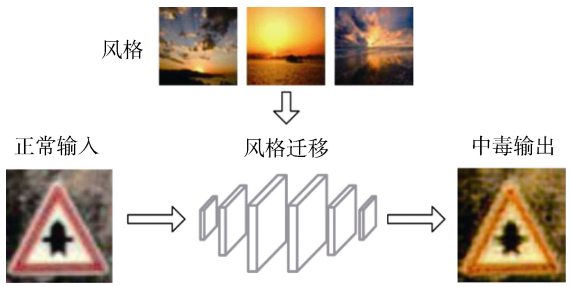


图 7 风格迁移后门攻击

Figure 7 Style transfer backdoor attack

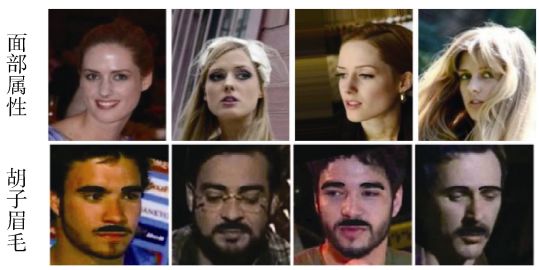


图 8 良性特征攻击

Figure 8 Benign feature attack

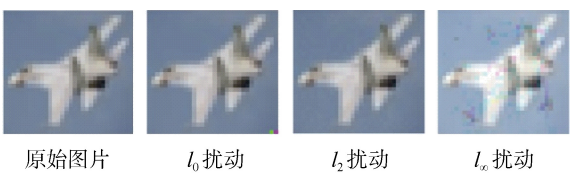


图 9 对抗扰动攻击

Figure 9 Adversarial perturbation attack

图 10 数字后门与物理后门

Figure 10 Digital backdoor and physical backdoor

不同于文献[17]中在图片中添加黑框眼镜或紫色太阳镜, Sarkar 等人[31]使用人脸图片中原有的一些面部的特定表情作为触发器, 例如微笑、眉毛挑起、眯起眼睛、嘴唇微张等, 成功在人脸识别模型中注入了后门, 这种面部表情也是来自真实世界的后门攻击。Xue 等人[28]则认为 FaceHack[31]中使用特定表情作为触发器, 会由于不经意的触发而被发现, 因此该方法将触发器隐藏到人脸的胡须和眉毛中, 可以更好的在真实场景中对人脸识别系统进行后门攻击。

受 DNN 模型可视化中颜色和边缘是主导特征的启发, Li 等人[59]提出了一种使用彩色条纹作为触发器的后门攻击方式, 该彩色条纹由特定调制波形的 LED 光产生。在现实场景中, 例如针对人脸识别系统, 只需在人脸图像采集处布置 LED, 进行后门攻击时 LED 闪烁即可为人脸图像注入触发器, 而 LED 的闪烁人类难以察觉, 十分隐蔽。并且该文在黑盒的条件下进行攻击, 无需访问原数据, 在其采集时通过 LED 添加触发器, 作为干净标签的中毒样本, 当管理方对模型进行重训练时就会植入后门。

4.2.12 特洛伊木马攻击

特洛伊木马攻击与之前基于数据投毒的方法不同, 其通常利用系统漏洞或恶意程序来篡改位于系统内存中的模型参数, 从而控制模型并植入后门。

Dumford 等人[19]提出了一种直接扰动模型参数的特洛伊木马攻击, 将后门注入的过程转换为对模型权重值的贪婪搜索过程。该文假设攻击者可以访问模型, 然后对模型权重应用不同的随机扰动, 每次仅扰动一层, 根据测试集选取最优权重, 然后继续扰动其他层, 最终得到注入了后门的中毒模型。

与文献[19]类似, Costales 等人[52]也提出了一种在模型运行时篡改内存中模型参数的攻击方法。攻击者可以使用恶意软件访问内存中的模型参数, 通过计算模型各神经元对于中毒样本的平均梯度, 选择少数具有较大梯度的参数进行篡改, 篡改数值根据模型对中毒数据集的重训练来确定。

Guo 等人[20]提出了一种特洛伊木马网络, 使模型在学习原任务的同时学习一个隐藏任务, 通过密钥编码一个特定的权重排列, 用于激活隐藏任务的模型参数, 从而激活模型后门。该文还证明了对于特洛伊木马网络的检测将是一个 NPC 问题, 因此在没有密钥的情况下隐藏任务是不可检测的。

Rakin 等人[21]提出了一种目标比特位木马 (TBT) 攻击方法, 首先通过神经梯度排序 (NRU) 确定与攻击目标相关联的比特位, 然后通过木马位搜索找出其中较为脆弱的比特位, 最后利用比特翻转攻击先前确定的比特位, 即可成功在模型中注入后门, 该方法相比于参数搜索的方法更加简单有效。

与前面修改模型参数的方法不同, Tang 等人[60]

提出了一种基于恶意子网络的特洛伊木马攻击, 其在模型中插入一个恶意后门模块, 当输入带有特殊触发器时恶意子网络会使模型错误分类到目标标签中, 且由于恶意模块仅与触发器有关而与模型无关, 这种方法适用于任意深度学习模型。

此外, Bagdasaryan 等人[49]提出了源代码中毒的后门攻击方式, 该方法不修改训练数据也不访问训练过程, 只尝试对源代码进行修改, 代码在训练过程中动态创建中毒输入, 而对于源代码的检测十分困难, 因此该种攻击具有很高的隐蔽性。

4.2.13 攻击方法总结

表 2 是对前述各后门攻击方法的总结, 其中敌手模型如 3.3 节所述, 表示了使用该方法所需的知识与能力。从表中可以看出, 绝大多数方法都依赖于修改原样本进行数据投毒, 部分方法利用模型参数来优化后门注入的过程。

表 2 深度学习中的后门攻击方法

Table 2 Backdoor attack methods in deep learning

攻击类型	攻击方法	敌手模型		相关实验					防御测试
		数据	模型	大小	形状	位置	透明度	投毒率	
简单攻击	BadNets[7]	✓	✗	✗	✓	✗	✗	✓	N/A
	Guo et al.[33]	✓	✗	—	—	—	—	✓	N/A
不可见攻击	Chen et al.[17]	✓	✗	✓	✗	✗	✓	✓	N/A
	Li et al.[27]	✓	✓	✓	✗	✗	—	✓	[42]
	WaNet[36]	✓	✗	✓	—	—	—	✓	[42], [60], [43], [53]
	Xue et al.[32]	✓	✗	—	—	—	—	✗	N/A
干净标签攻击	Turner et al.[23]	✓	✓	✗	✗	✓	✓	✓	N/A
	M.Barni et al.[24]	✓	✗	✗	✗	✗	✗	✓	N/A
	Saha et al.[25]	✓	✓	✓	✗	✓	✗	✓	[60]
	Quiring et al.[39]	✓	✗	✗	✗	✗	—	✓	[62]
特定标签攻击	Ning et al.[40]	✓	✗	✓	✗	✗	✗	✓	N/A
	Li et al.[41]	✓	✗	✗	✗	✗	✓	✓	[42], [43], [53], [54]
反向工程	Liu et al.[18]	✗	✓	✓	✓	✗	✓	✗	N/A
	Yao et al.[46]	✓	✓	✗	✗	✗	✗	✓	[42], [43]
多后门攻击	Cheng et al.[22]	✓	✓	—	✗	—	—	✓	[42], [63], [48]
	Xue et al.[47]	✓	✗	✗	✗	✓	✓	✓	[42], [64]
良性特征攻击	Lin et al.[32]	✓	✗	✓	✓	✓	✗	✓	[43], [48]
	FaceHack[28]	✓	✗	✓	✓	✓	—	✓	[42], [61], [64], [53] [65], [48], [26]
对抗扰动攻击	Liao et al.[34]	✓	✗	—	—	—	—	✓	N/A
	Garg et al.[50]	✓	✓	✗	✗	✗	✗	✗	N/A
适应性攻击	Zhao et al.[38]	✓	✓	✓	✗	✓	—	✓	[42], [60]
	Tan et al.[51]	✓	✓	✗	✗	✗	✗	✓	[42], [60], [64]
	Vitaly et al.[49]	✗	✗	✓	✗	✗	✗	✓	[42], [54], [55]
	Ali et al.[56]	✓	✗	✗	✗	✗	✗	✗	[53], [55], [63], [66]
动态触发攻击	Nguyen et al.[29]	✓	✗	—	—	—	—	✗	[42], [43], [67]
	Salem et al.[30]	✓	✗	✓	—	✓	✓	✓	[42], [53], [48]
真实物理攻击	Li et al.[59]	✗	✗	—	—	—	—	✓	[42], [68], [69]
	Refool[57]	✓	✗	—	—	—	—	✓	[42], [43]
特洛伊木马攻击	Dumford et al.[19]	✓	✓	—	—	—	—	✓	N/A
	Costales et al.[52]	✓	✓	✗	✗	✗	✗	✓	[53]
	Tang et al.[60]	✗	✓	✗	✗	✗	✗	✗	[42], [70]
	TBT[21]	✓	✓	✗	✗	✓	✓	✓	N/A
	TrojanNet[20]	✓	✓	—	—	—	—	✗	N/A

(注: —在本文中表示某属性不适用于该方法, N/A 表示缺失。)

如 4.1.1 节所述, 触发器各属性以及投毒率都会对于后门植入的过程与性能具有一定程度的影响, 许多攻击算法会对其进行实验分析, 因此本文在表 2 中相关实验部分对其进行了总结, 表示该方法是否对该属性进行了鲁棒性实验, 其中投毒率表示注入的中毒数据占全部数据集的比例, 在一些篡改模型参数的攻击中也表示篡改参数所占比例。

近年来也出现了许多优秀后门防御算法, 例如基于触发器特征与良性特征差异进行检测的 Neural Cleanse [42] 算法、基于样本熵分布差异进行检测的 Strip[53] 算法、基于模型内部神经元反应差异进行检测的 ABS[48] 算法等等, 因此本文在表 2 中的防御测试部分总结了各攻击方法针对相关防御方法进行的攻击测试实验。从表中可以看出, 适应性攻击由于其对防御算法的针对性设计, 可以攻破大多数防御算法, 特定标签攻击仅与特定类样本建立联系, 使攻击更加隐蔽, 可以躲避大多数检测不同样本差异的防御手段。良性特征攻击不添加新的特征, 同样可以躲过基于样本特征差异进行检测的方法。

4.3 其他领域的后门攻击

后门攻击最早基于图像数据提出, 因此后续的大多数研究都针对以图像为数据载体的计算机视觉领域进行。近年来, 后门攻击开始越来越多的涉及其他领域的任务, 不同领域和不同任务之间的后门攻击可以相互借鉴但又存在着巨大的差异。

其他领域后门攻击的研究方向与计算机视觉领域类似, 同样围绕可见性、触发器类型、标签类型、攻击形式、场景设置等方向进行研究, 其不同之处主要体现在数据载体和学习范式。以自然语言处理为例, 其以文本为数据载体, 与图像像素的连续空间不同, 文本进行词嵌入后的数据空间是离散的, 这对触发器的设计提出了新的要求, 同时为了保证隐匿性, 如何对字词进行增删改操作的同时保证句子的流畅性也是新的难点。

Dai 等人[71]探索了基于 LSTM 模型的文本分类任务的后门攻击, 以特定语句作为触发器, 使模型误分类为目标标签。Kurita 等人[72]则针对自然语言处理领域最近活跃的大量预训练模型, 以某些关键字作为触发器使模型权重中毒, 用户下载预训练模型进行微调后便被植入了后门。Chan 等人[73]使用条件对抗正则化的自编码器基于潜空间特征生成中毒句子对, 保证了中毒句子仍然连贯且符合语法。Chen 等人[74]则进一步对自然语言处理领域的后门攻击进行了归纳, 将触发机制分为了三个层次, 分别为字符层次、词层次、句子层次, 并表明一个成功的触发器应当不改变原句的正常标签, 而是误导模型分类为目标标签。

迁移学习也是后门攻击的热门领域, Yao 等人[46]首先探索了迁移学习中的后门攻击, 其基于反向工程生成与冻结层相关联的触发器, 从而在用户迁移学习时保留后门, 随后出现了一系列相关研究[75-76]。Yan 等人[77]针对半监督学习, 基于对抗扰动生成中毒的未标签数据, 在半监督模型中注入后门。Zhai 等人[78]针对语音识别领域的说话人鉴别, 通过聚类构造中毒数据来感染模型。此外, 还有许多后门攻击的相关研究针对联邦学习[79-82]、强化学习[83-84]、图神经网络[85-86]等领域进行, 后门攻击已经在越来越多的领域和任务上产生了严重的威胁。

5 常用数据集及模型总结

该节针对深度学习后门攻击中常用的数据集与深度学习模型, 按照任务进行了分类总结, 数据集总结如表 3 所示, 模型总结如表 4 所示, 两表中删去了只出现一次的数据集或模型。

常用数据集和模型主要集中在分类任务上, 计算机视觉领域多从手写数字识别、人脸识别、目标识别以及交通信号识别任务进行攻击, 自然语言处理领域则主要集中在文本分类以及情感分析任务上。除此之外还有一些针对语音识别[18]、虹膜识别[46]、目标检测[32]以及恶意文件检测[52]等任务的攻击。

6 后门攻击的有益应用

后门攻击作为一种方法, 并非完全只能用于攻击模型来产生安全威胁, 不少学者开始研究使用后门攻击的方法反过来保护模型和数据。

Adi 等人[102]在一些用于商业的模型中注入后门, 作为一种跟踪机制以保护模型的知识产权, 称为模型水印, 该水印对于模型的正常功能没有显著影响。具体的, 该方法基于密钥生成水印, 然后利用后门攻击的方法将水印注入到模型中, 并给出了对于模型水印的验证方法。Li 等人[103]针对开源数据集的保护问题, 提出了一种基于后门嵌入的数据集水印方法, 使用后门攻击中构造中毒数据集的方法, 以触发器作为水印并给定目标标签, 通过验证第三方模型是否存在对应后门来确定该数据集是否用于训练模型。

此外, 也有学者使用后门来探索深度学习模型的可解释性, Lin 等人[104]使用后门来量化评估一个人工智能方法的可解释性, 通过检验一个人工智能方

表 3 后门攻击中的常用数据集

Table 3 Common datasets in backdoor attacks

任务	数据集	相关文献
手写数字识别	Mnist[87]	[7], [24], [46], [47], [27], [29], [34], [52], [49], [30], [36], [37], [19]
目标识别	CIFAR-10[88]	[23], [47], [27], [32], [29], [25], [51], [34], [50], [52], [39], [30], [36], [22], [37], [21], [20]
	SVHN[89]	[21], [20]
	ImageNet[90]	[25], [57], [41], [49], [22], [37], [21], [71]
人脸识别	VGG Face[91]	[17], [18], [46], [22]
	VGG Face2[92]	[31], [59], [33]
	LFW[93]	[18], [32], [59], [33]
	YouTube Aligned Face[94]	[17], [47], [32], [59], [33], [28], [71]
	PubFig[95]	[46], [57], [59], [30]
	CelebA[96]	[31], [30], [36]
交通信号识别	GTSRB[97]	[24], [46], [27], [32], [29], [25], [51], [34], [36], [22], [37], [20], [71]
情感分析	IMDb Movie Reviews[98]	[49], [71], [72], [74]
	Yelp Review[99]	[72], [73]
	Amazon Reviews[100]	[72], [74]
	Stanford Sentiment Treebank[101]	[72], [74]

表 4 后门攻击中的常用模型

Table 4 Common models in backdoor attacks

任务	模型	相关文献
手写数字识别	简单卷积网络	[7], [24], [46], [27], [29], [52], [49], [30], [36], [19]
	LeNet[105]	[47], [34]
目标识别	简单卷积网络	[32], [39]
	VGG[106]	[47], [51], [34], [30], [22], [21]
	ResNet[107]	[23], [27], [29], [57], [50], [41], [49], [36], [22], [21], [20]
	DenseNet[108]	[51], [57]
人脸识别	VGG-Face[91]	[17], [18], [46], [47], [32], [59], [22], [28]
	DeepID[109]	[17], [28]
	ResNet	[57], [41], [31], [21], [22], [19]
	Inception[110]	[31], [33]
交通信号识别	简单卷积网络	[46], [32]
	LeNet	[24], [34]
	VGG	[51], [22]
	ResNet	[27], [29], [57], [36], [22], [20]
情感分析	LSTM[111]	[71]
	XLNET[112]	[72], [73]
	BERT[113]	[72], [73]
	RoBERT[114]	[49], [73]

法是否能够检测到输入中存在的后门触发器来评估其可解释性, 该文认为触发器导致的错误分类是一个基本事实, 而一个鲁棒的可解释性人工智能方法应当可以识别出与该预测相关联的真正区域。

7 未来研究方向

在前文中已经全面的分析了近年来深度学习后门攻击的相关研究, 从中可以看出, 目前仍然存在

很多关键的问题需要解决, 这一节对未来后门攻击研究的关键方向进行了展望。

1) 更多应用: 后门攻击作为一种方法, 不仅仅是产生安全威胁, 也可以在其他方向上发挥作用。目前已经出现了一些有益应用, 但仍然还存在很多针对不同领域的潜在应用。例如, 可以将触发器作为一种授权模型使用的密钥, 只有输入带有触发器时才可正常使用。

2) 更加真实: 目前针对黑盒模型和真实物理世界的后门攻击研究较少, 这种攻击具有更大的威胁, 也更具有研究意义。使后门攻击方法在黑盒模型的条件下更具隐匿性、更好的抵抗来自物理世界可能产生的干扰是未来的一大挑战。

3) 触发器设计: 目前触发器的研究主要针对其大小、形状、位置以及不可见性, 而针对其潜在特征表示的深入研究较少, 同时缺少跨领域的触发器通用设计方法, 因此如何更好的设计触发器也将是未来的重点研究方向。

4) 可解释性: 目前后门攻击仅依据实验效果, 而没有完整有效的理论支撑, 什么样的模型更容易嵌入后门, 什么样的触发器更容易被模型学习, 相关的可解释性讨论与分析也是一大研究方向。

8 结论

本文针对深度学习中现有的后门攻击方法进行了全面的分析与归类, 并对常用数据集和模型进行了总结。目前对于深度学习模型的后门攻击仍不成熟, 存在很多可研究的方向, 是一个深度学习中正在飞速发展的领域, 希望本文可以为今后的后门攻击研究提供一个总结性的参考。

参考文献

[1] Adjabi I, Ouahabi A, Benzaoui A, et al. Past, Present, and Future of Face Recognition: A Review[J]. *Electronics*, 2020, 9(8): 1188.

[2] Baomar H, Bentley P J. An Intelligent Autopilot System that Learns Piloting Skills from Human Pilots by Imitation[C]. *2016 International Conference on Unmanned Aircraft Systems*, 2016: 1023-1031.

[3] Bahdanau D, Cho K, Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate[J]. *CoRR*, 2014, abs/1409.0473.

[4] Graves A, Mohamed A R, Hinton G. Speech Recognition with Deep Recurrent Neural Networks[C]. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013: 6645-6649.

[5] Brown T B, Mann B, Ryder N, et al. Language models are few-shot learners[EB/OL]. 2020: ArXiv Preprint ArXiv:2005.14165.

[6] Shokri R, Stronati M, Song C Z, et al. Membership Inference Attacks Against Machine Learning Models[C]. *2017 IEEE Symposium on Security and Privacy*, 2017: 3-18.

[7] Gu T, Dolan-Gavitt B, Garg S. Badnets: Identifying vulnerabilities in the machine learning model supply chain[EB/OL]. 2017: ArXiv Preprint ArXiv:1708.06733.

[8] Goodfellow I J, Shlens J, Szegedy C. Explaining and Harnessing Adversarial Examples[J]. *CoRR*, 2014, abs/1412.6572.

[9] Madry A, Makelov A, Schmidt L, et al. Towards deep learning models resistant to adversarial attacks[EB/OL]. 2017: ArXiv Preprint ArXiv:1706.06083.

[10] Xu J, Li Y M, Jiang Y, et al. Adversarial Defense via Local Flatness Regularization[C]. *2020 IEEE International Conference on Image Processing*, 2020: 2196-2200.

[11] Fan Y B, Wu B Y, Li T H, et al. Sparse Adversarial Attack via Perturbation Factorization[M]. Computer Vision – ECCV 2020. Cham: Springer International Publishing, 2020: 35-50.

[12] Barreno M, Nelson B, Sears R, et al. Can Machine Learning Be Secure? [C]. *The 2006 ACM Symposium on Information, computer and communications security*, 2006: 16-25.

[13] Biggio B, Nelson B, Laskov P. Support Vector Machines under Adversarial Label Noise[C]. *Asian conference on machine learning*, 2011: 97-112.

[14] M. Kloft, P. Laskov. Online anomaly detection under adversarial impact[C]. *The Thirteenth International Conference on Artificial Intelligence and Statistics*, 2010: 405-412.

[15] A. Shafahi, R. Huang, M. Najibi, et al. Poison frogs! Targeted clean-label poisoning attacks on neural networks[C]. *Advances in Neural Information Processing Systems*, 2018: 6103-6113.

[16] Li Y, Wu B, Jiang Y, et al. Backdoor learning: A survey[EB/OL]. 2020: ArXiv Preprint ArXiv:2007.08745.

[17] Chen X Y, Liu C, Li B, et al. Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning[EB/OL]. 2017: arXiv: 1712.05526[cs.CR]. https://arxiv.org/abs/1712.05526.

[18] Liu Y Q, Ma S Q, Aafer Y, et al. Trojaning Attack on Neural Networks[C]. *The 2018 Network and Distributed System Security Symposium*, 2018: 1-15.

[19] Dumford J, Scheirer W. Backdooring Convolutional Neural Networks via Targeted Weight Perturbations[C]. *2020 IEEE International Joint Conference on Biometrics*, 2020: 1-9.

[20] Guo C, Wu R, Weinberger K Q. Trojannet: Embedding hidden trojan horse models in neural networks[EB/OL]. 2020: ArXiv Preprint ArXiv:2002.10078.

[21] Rakin A S, He Z Z, Fan D L. TBT: Targeted Neural Network Attack with Bit Trojan[C]. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020: 13195-13204.

[22] Cheng S, Liu Y, Ma S, et al. Deep Feature Space Trojan Attack of Neural Networks by Controlled Detoxification[J]. *The Association for the Advance of Artificial Intelligence*, 2021: 1148-1156.

[23] Turner A, Tsipras D, Madry A. Label-consistent backdoor attacks[EB/OL]. 2019: ArXiv Preprint ArXiv:1912.02771.

[24] Barni M, Kallas K, Tondi B. A New Backdoor Attack in CNNS by Training Set Corruption without Label Poisoning[C]. *2019 IEEE*

International Conference on Image Processing, 2019: 101-105.

[25] Saha A, Subramanya A, Pirsiavash H. Hidden Trigger Backdoor Attacks[J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, 34(7): 11957-11965.

[26] Veldanda A K, Liu K, Tan B, et al. NNoculation: broad spectrum and targeted treatment of backdoored DNNs[EB/OL]. 2020: ArXiv Preprint ArXiv:2002.08313.

[27] Li S F, Xue M H, Zhao B Z H, et al. Invisible Backdoor Attacks on Deep Neural Networks via Steganography and Regularization[J]. *IEEE Transactions on Dependable and Secure Computing*, 2021, 18(5): 2088-2105.

[28] Xue M F, He C, Wang J, et al. Backdoors Hidden in Facial Features: A Novel Invisible Backdoor Attack Against Face Recognition Systems[J]. *Peer-to-Peer Networking and Applications*, 2021, 14(3): 1458-1474.

[29] Nguyen A, Tran A. Input-Aware Dynamic Backdoor Attack[EB/OL]. 2020: ArXiv Preprint ArXiv:2010.08138.

[30] Salem A, Wen R, Backes M, et al. Dynamic backdoor attacks against machine learning models[EB/OL]. 2020: ArXiv Preprint ArXiv:2003.03675.

[31] Sarkar E, Benkraouda H, Maniatakos M. FaceHack: Triggering Backdoored Facial Recognition Systems Using Facial Characteristics[EB/OL]. 2020: ArXiv Preprint ArXiv:2006.11623.

[32] Lin J Y, Xu L, Liu Y Q, et al. Composite Backdoor Attack for Deep Neural Network by Mixing Existing Benign Features[C]. *The 2020 ACM SIGSAC Conference on Computer and Communications Security*, 2020: 113-131.

[33] Guo W, Tondi B, Barni M. A Master Key Backdoor for Universal Impersonation Attack Against DNN-Based Face Verification[J]. *Pattern Recognition Letters*, 2021, 144: 61-67.

[34] Zhong H T, Liao C, Squicciarini A C, et al. Backdoor Embedding in Convolutional Neural Network Models via Invisible Perturbation[EB/OL]. 2018: ArXiv Preprint ArXiv:1808.10307.

[35] Moosavi-Dezfooli S M, Fawzi A, Fawzi O, et al. Universal Adversarial Perturbations[C]. *2017 IEEE Conference on Computer Vision and Pattern Recognition*, 2017: 86-94.

[36] Nguyen A, Tran A. WaNet - Imperceptible Warping-Based Backdoor Attack[EB/OL]. 2021: ArXiv Preprint ArXiv:2102.10369.

[37] Ning R, Li J, Xin C S, et al. Invisible Poison: A Blackbox Clean Label Backdoor Attack to Deep Neural Networks[C]. *IEEE INFOCOM 2021 - IEEE Conference on Computer Communications*, 2021: 1-10.

[38] Zhao S H, Ma X J, Zheng X, et al. Clean-Label Backdoor Attacks on Video Recognition Models[C]. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020: 14431-14440.

[39] Quiring E, Rieck K. Backdooring and Poisoning Neural Networks with Image-Scaling Attacks[C]. *2020 IEEE Security and Privacy Workshops*, 2020: 41-47.

[40] Xiao Q, Chen Y, Shen C, et al. Seeing is not believing: Camouflage attacks on image scaling algorithms[C]. *28th USENIX Security Symposium*, 2019: 443-460.

[41] Li Y, Li Y, Wu B, et al. Backdoor Attack with Sample-Specific Triggers[EB/OL]. 2020: ArXiv Preprint ArXiv:2012.03816.

[42] Wang B L, Yao Y S, Shan S, et al. Neural Cleanse: Identifying and Mitigating Backdoor Attacks in Neural Networks[C]. *2019 IEEE Symposium on Security and Privacy*, 2019: 707-723.

[43] Liu K, Dolan-Gavitt B, Garg S. Fine-Pruning: Defending Against Backdooring Attacks on Deep Neural Networks[C]. *Research in Attacks*, *Intrusions*, *and Defenses*, 2018: 273-294.

[44] Tancik M, Mildenhall B, Ng R. StegaStamp: Invisible Hyperlinks in Physical Photographs[C]. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020: 2114-2123.

[45] Baluja S. Hiding images in plain sight: Deep steganography[C]. *The 31st International Conference on Neural Information Processing Systems*, 2017: 2066-2076.

[46] Yao Y S, Li H Y, Zheng H T, et al. Latent Backdoor Attacks on Deep Neural Networks[C]. *The 2019 ACM SIGSAC Conference on Computer and Communications Security*, 2019: 2041-2055.

[47] Xue M F, He C, Wang J, et al. One-to-N & N-to-One: Two Advanced Backdoor Attacks Against Deep Learning Models[J]. *IEEE Transactions on Dependable and Secure Computing*, 8448, PP(99): 1.

[48] Liu Y Q, Lee W C, Tao G H, et al. ABS: Scanning Neural Networks for Back-Doors by Artificial Brain Stimulation[C]. *The 2019 ACM SIGSAC Conference on Computer and Communications Security*, 2019: 1265-1282.

[49] Bagdasaryan E, Shmatikov V. Blind backdoors in deep learning models[EB/OL]. 2020: ArXiv Preprint ArXiv:2005.03823.

[50] Garg S, Kumar A, Goel V, et al. Can Adversarial Weight Perturbations Inject Neural Backdoors[C]. *The 29th ACM International Conference on Information & Knowledge Management*, 2020: 2029-2032.

[51] Tan T J L, Shokri R. Bypassing Backdoor Detection Algorithms in Deep Learning[C]. *2020 IEEE European Symposium on Security and Privacy*, 2020: 175-183.

[52] Costales R, Mao C Z, Norwitz R, et al. Live Trojan Attacks on Deep Neural Networks[C]. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020: 3460-3469.

[53] Gao Y S, Xu C G, Wang D R, et al. STRIP: A Defence Against Trojan Attacks on Deep Neural Networks[C]. *The 35th Annual Computer Security Applications Conference*, 2019: 113-125.

[54] Chou E, Tramèr F, Pellegrino G. SentiNet: Detecting Localized Universal Attacks Against Deep Learning Systems[C]. *2020 IEEE Security and Privacy Workshops*, 2020: 48-54.

[55] Hong S, Chandrasekaran V, Kaya Y, et al. On the effectiveness of mitigating data poisoning attacks with gradient shaping[EB/OL]. 2020: ArXiv Preprint ArXiv:2002.11497.

[56] Ali H, Nepal S, Kanhere S S, et al. HaS-Nets: A Heal and Select Mechanism to Defend DNNs Against Backdoor Attacks for Data Collection Scenarios[EB/OL]. 2020: ArXiv Preprint ArXiv:2012.07474.

[57] Liu Y F, Ma X J, Bailey J, et al. Reflection Backdoor: A Natural Backdoor Attack on Deep Neural Networks[C]. *Computer Vision – ECCV 2020*, 2020: 182-199.

[58] Hendrycks D, Zhao K, Basart S, et al. Natural Adversarial Examples[C]. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021: 15257-15266.

[59] Li H, Wang Y, Xie X, et al. Light Can Hack Your Face! Black-box

Backdoor Attack on Face Recognition Systems[EB/OL]. 2020: ArXiv Preprint ArXiv:2009.06996.

[60] Tang R X, Du M N, Liu N H, et al. An Embarrassingly Simple Approach for Trojan Attack in Deep Neural Networks[C]. *The 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020: 218-228.

[61] Tran B, Li J, Madry A. Spectral signatures in backdoor attacks[EB/OL]. 2018: ArXiv Preprint ArXiv:1811.00636.

[62] Quiring E, Klein D, Arp D, et al. Adversarial preprocessing: Understanding and preventing image-scaling attacks in machine learning[C]. *29th USENIX Security Symposium*, 2020: 1363-1380.

[63] Kolouri S, Saha A, Pirsiavash H, et al. Universal Litmus Patterns: Revealing Backdoor Attacks in CNNS[C]. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020: 298-307.

[64] Chen B, Carvalho W, Baracaldo N, et al. Detecting backdoor attacks on deep neural networks by activation clustering[EB/OL]. 2018: ArXiv Preprint ArXiv:1811.03728.

[65] Sarkar E, Alkindi Y, Maniatakos M. Backdoor Suppression in Neural Networks Using Input Fuzzing and Majority Voting[J]. *IEEE Design & Test*, 2020, 37(2): 103-110.

[66] Doan B G, Abbasnejad E, Ranasinghe D C. Februus: Input Purification Defense Against Trojan Attacks on Deep Neural Network Systems[C]. *ACSAC '20: Annual Computer Security Applications Conference*, 2020: 897-912.

[67] Zhao P, Chen P Y, Das P, et al. Bridging mode connectivity in loss landscapes and adversarial robustness[EB/OL]. 2020: ArXiv Preprint ArXiv:2005.00060.

[68] Cheng H, Xu K, Liu S, et al. Defending against backdoor attack on deep neural networks[EB/OL]. 2020: ArXiv Preprint ArXiv: 2002.12162.

[69] Wang J L, Huang T Z, Zhao X L, et al. Reweighted Block Sparsity Regularization for Remote Sensing Images Destriping[J]. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2019, 12(12): 4951-4963.

[70] Huang X J, Alzantot M, Srivastava M. NeuronInspect: Detecting Backdoors in Neural Networks via Output Explanations[EB/OL]. 2019: ArXiv Preprint ArXiv:1911.07399.

[71] Dai J Z, Chen C S, Li Y F. A Backdoor Attack Against LSTM-Based Text Classification Systems[J]. *IEEE Access*, 2019, 7: 138872-138878.

[72] Kurita K, Michel P, Neubig G. Weight poisoning attacks on pre-trained models[EB/OL]. 2020: ArXiv Preprint ArXiv:2004. 06660.

[73] Chan A, Tay Y, Ong Y S, et al. Poison Attacks Against Text Datasets with Conditional Adversarially Regularized Autoencoder[EB/OL]. 2020: ArXiv Preprint ArXiv:2010.02684.

[74] Chen X, Salem A, Backes M, et al. Badnl: Backdoor attacks against nlp models[EB/OL]. 2020: ArXiv Preprint ArXiv: 2006.01043.

[75] Wang S, Nepal S, Rudolph C, et al. Backdoor Attacks Against Transfer Learning with Pre-Trained Deep Learning Models[J]. *IEEE Transactions on Services Computing*, 2020, PP(99): 1.

[76] Zhang Z, Xiao G, Li Y, et al. Red Alarm for Pre-trained Models: Universal Vulnerabilities by Neuron-Level Backdoor Attacks[EB/OL]. 2021: ArXiv Preprint ArXiv:2101.06969.

[77] Zhicong Yan1, Gaolei Li1, Yuan Tian, et al. DeHiB: Deep Hidden Backdoor Attack on Semi-supervised Learning via Adversarial Perturbation[C]. *The Association for the Advance of Artificial Intelligence*, 2021.

[78] Zhai T Q, Li Y M, Zhang Z Q, et al. Backdoor Attack Against Speaker Verification[C]. *ICASSP 2021 - 2021 IEEE International Conference on Acoustics*, *Speech and Signal Processing*, 2021: 2560-2564.

[79] Bagdasaryan E, Veit A, Hua Y, et al. How to backdoor federated learning[C]. *International Conference on Artificial Intelligence and Statistics*. 2020: 2938-2948.

[80] Sun Z, Kairouz P, Suresh A T, et al. Can you really backdoor federated learning? [EB/OL]. 2019: ArXiv Preprint ArXiv:1911. 07963.

[81] Wang H, Sreenivasan K, Rajput S, et al. Attack of the tails: Yes, you really can backdoor federated learning[EB/OL]. 2020: ArXiv Preprint ArXiv:2007.05084.

[82] Chen C L, Golubchik L, Paolieri M. Backdoor attacks on federated meta-learning[EB/OL]. 2020: ArXiv Preprint ArXiv:2006.07026.

[83] Kiourti P, Wardega K, Jha S, et al. TrojDRL: Evaluation of Backdoor Attacks on Deep Reinforcement Learning[C]. *2020 57th ACM/IEEE Design Automation Conference*, 2020: 1-6.

[84] Wang Y, Sarkar E, Maniatakos M, et al. Stop-and-Go: Exploring Backdoor Attacks on Deep Reinforcement Learning-Based Traffic Congestion Control Systems[EB/OL]. 2020: arXiv: 2003. 07859[cs.CR]. https://arxiv.org/abs/2003.07859.

[85] Xi Z, Pang R, Ji S, et al. Graph backdoor[EB/OL]. 2020: ArXiv Preprint ArXiv:2006.11890.

[86] Zhang Z X, Jia J Y, Wang B H, et al. Backdoor Attacks to Graph Neural Networks [EB/OL]. 2020: ArXiv Preprint ArXiv:2006. 11165.

[87] Deng L. The MNIST Database of Handwritten Digit Images for Machine Learning Research [Best of the Web[J]. *IEEE Signal Processing Magazine*, 2012, 29(6): 141-142.

[88] Krizhevsky A, Hinton G. Learning multiple layers of features from tiny images. Technical report, 2009.

[89] Netzer Y, Wang T, Coates A, et al. Reading digits in natural images with unsupervised feature learning[J]. NIPS Workshop on Deep Learning and Unsupervised Feature Learning, 2011.

[90] Deng J, Dong W, Socher R, et al. ImageNet: A Large-Scale Hierarchical Image Database[C]. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009: 248-255.

[91] Parkhi O M, Vedaldi A, Zisserman A. Deep Face Recognition[C]. *The British Machine Vision Conference 2015*, 2015.

[92] Cao Q, Shen L, Xie W D, et al. VGGFace2: A Dataset for Recognising Faces across Pose and Age[C]. *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition*, 2018: 67-74.

[93] Huang G B, Mattar M, Berg T, et al. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments[C]. *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.

[94] Wolf L, Hassner T, Maoz I. Face Recognition in Unconstrained Videos with Matched Background Similarity[C]. *CVPR 2011*, 2011: 529-534.

[95] Kumar N, Berg A C, Belhumeur P N, et al. Attribute and Simile Classifiers for Face Verification[C]. *2009 IEEE 12th International Conference on Computer Vision*, 2009: 365-372.

[96] Liu Z W, Luo P, Wang X G, et al. Deep Learning Face Attributes in the Wild[C]. *2015 IEEE International Conference on Computer Vision*, 2015: 3730-3738.

[97] Stallkamp J, Schlipsing M, Salmen J, et al. The German Traffic Sign Recognition Benchmark: A Multi-Class Classification Competition[C]. *The 2011 International Joint Conference on Neural Networks*, 2011: 1453-1460.

[98] Maas A, Daly R E, Pham P T, et al. Learning word vectors for sentiment analysis[C]. *The 49th annual meeting of the association for computational linguistics: Human language technologies*, 2011: 142-150.

[99] Zhang X, Zhao J J, LeCun Y. Character-Level Convolutional Networks for Text Classification[J]. *CoRR*, 2015, abs/1509.01626.

[100] Blitzer J, Dredze M, Pereira F. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification[C]. *The 45th annual meeting of the association of computational linguistics*, 2007: 440-447.

[101] Socher R, Perelygin A, Wu J, et al. Recursive deep models for semantic compositionality over a sentiment treebank[C]. *The 2013 conference on empirical methods in natural language processing*, 2013: 1631-1642.

[102] Adi Y, Baum C, Cisse M, et al. Turning your weakness into a strength: Watermarking deep neural networks by backdooring[C]. *27th USENIX Security Symposium*, 2018: 1615-1631.

[103] Li Y, Zhang Z, Bai J, et al. Open-sourced Dataset Protection via Backdoor Watermarking[EB/OL]. 2020: ArXiv Preprint ArXiv: 2010.05821.

[104] Lin Y S, Lee W C, Celik Z B. What do You See? : Evaluation of Explainable Artificial Intelligence (XAI) Interpretability through Neural Backdoors[EB/OL].2020: ArXiv Preprint ArXiv:2009. 10639.

[105] Lecun Y, Bottou L, Bengio Y, et al. Gradient-Based Learning Applied to Document Recognition[J]. *Proceedings of the IEEE*, 1998, 86(11): 2278-2324.

[106] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition[J]. *CoRR*, 2014, abs/1409.1556.

[107] He K M, Zhang X Y, Ren S Q, et al. Deep Residual Learning for Image Recognition[C]. *2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016: 770-778.

[108] Huang G, Liu Z, van der Maaten L, et al. Densely Connected Convolutional Networks[C]. *2017 IEEE Conference on Computer Vision and Pattern Recognition*, 2017: 2261-2269.

[109] Sun Y, Wang X G, Tang X O. Deep Learning Face Representation from Predicting 10, 000 Classes[C]. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014: 1891-1898.

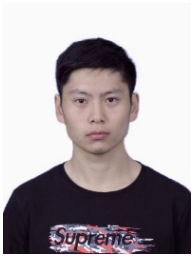
[110] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the Inception Architecture for Computer Vision[C]. *2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016: 2818-2826.

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

[112] Yang Z L, Dai Z H, Yang Y M, et al. XLNet: Generalized Autoregressive Pretraining for Language Understanding[J]. *CoRR*, 2019, abs/1906.08237.

[113] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[EB/OL]. 2018: ArXiv Preprint ArXiv:1810.04805.

[114] Liu Y H, Ott M, Goyal N, et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach[J]. *CoRR*, 2019, abs/1907.11692.



杜巍 于 2020 年在西安电子科技大学电子信息工程专业获得学士学位。现在上海交通大学网络空间安全专业攻读博士学位。研究领域为自然语言处理、人工智能安全。研究兴趣包括: 自然语言处理、人工智能安全。Email: dddddw@sjtu.edu.cn

刘功申 于 2003 年在上海交通大学计算机专业获得博士学位。现任上海交通大学网络空间安全学院教授。研究领域为人工智能安全、自然语言处理。研究兴趣包括: 人工智能安全、自然语言理解、内容安全、恶意代码防范等。Email: lgshen@sjtu.edu.cn