中图法分类号:TP309; TP18　　**文献标识码**: A　　**文章编号**: 1006-8961(2026)01-0045-17

论文引用格式:Long L H, Wang Z C and Zhang X P. 2026. Overview of neural network model steganography. Journal of Image and Graphics, 31(1): 0045-0061(龙玲慧, 王子驰, 张新鹏. 2026. 神经网络模型隐写研究进展. 中国图象图形学报, 31(1):0045-0061)[DOI:10.11834/jig.250267]

神经网络模型隐写研究进展

龙玲慧,王子驰*,张新鹏

上海大学通信与信息工程学院,上海　200444

摘　要: 神经网络模型数量增长迅猛,以神经网络为代表的人工智能技术在很多应用领域取得巨大成功。与此同时,神经网络模型含有大量冗余信息,可为隐藏机密信息提供便利条件,因此可以借助神经网络模型传递机密信息。在此背景下,本文介绍以神经网络为载体的隐写技术。通过与相关技术进行对比,首先概述了神经网络模型隐写的研究意义、基础概念和评价指标;之后依据模型隐写的不同策略,从基于训练的模型隐写、基于修改的模型隐写、基于后门等技术的模型隐写3个不同的角度分别梳理了研究现状,阐述各类方法的核心机制与适用场景,以及分析了各类方法在实际应用中的优缺点。同时也对模型隐写分析的成果进行了分析和讨论,总结白盒和黑盒模型隐写分析技术,揭示当前模型隐写攻防态势。最后对模型隐写技术发展趋势进行了展望,指出大模型隐写、高隐蔽—大容量协同优化、端到端安全传输等未来方向。本文提供了一个关于模型隐写技术的全面视角,旨在展示其在信息安全领域的重要性和潜力。

关键词: 隐写;模型隐写;隐写分析;神经网络;信息隐藏

Overview of neural network model steganography

Long Linghui, Wang Zichi*, Zhang Xinpeng

School of Communication and Information Engineering, Shanghai University, Shanghai　200444, China

Abstract: In recent years, the number of neural network models has increased rapidly, and artificial intelligence technology represented by neural networks has achieved great success in many application fields. Neural network models inherently contain considerable redundant information. This redundancy creates favorable conditions for hiding confidential data. Therefore, neural network models can be used as covers for covert communication. This new paradigm is called neural network model steganography (model steganography). The steganographer chooses the location where confidential information is embedded in the model and uses a key to embed the confidential information into the model for transmission. The receiver uses the shared key to extract the confidential information in the location where it is embedded. Model steganography is used for covert communication without detection. In recent years, neural network model steganography technology has made great progress. In practice, it can be applied in some scenarios, such as military defense or secret communication between intelligence agencies, embedding confidential information in the model training process or hiding secret tasks in the model. In command distribution, the commander intends to send different commands to multiple officers, or multiple officers send different messages to the commander. Using model steganography allows transferring confidential information without being detected. Meanwhile, by modifying model parameters, malicious developers can embed malicious

收稿日期:2025-06-19;**修回日期**:2025-08-01;**预印本日期**:2025-08-08

*通信作者:王子驰　wangzichi@shu.edu.cn

基金项目:国家自然科学基金项目(62376148)

Supported by:National Natural Science Foundation of China(62376148)

software into the benign model, resulting in the loss of model users. Using neural network backdoor technology to poison the target model enables performing different tasks defined by the attackers without the users' knowledge. Technologies related to model steganography include model watermarking and multimedia steganography based on a neural network model. The model watermark takes the neural network as the protection object and embeds the digital watermark in the model to protect the intellectual property rights of the model owner. The watermark information embedded in the model can be extracted correctly without affecting the normal use of the cover model and without deliberately concealing the existence of the watermark information. In addition, the embedding capacity can accommodate the watermark information, so there is no need to pursue large capacity. Multimedia steganography based on a neural network model takes multimedia data as covers and the neural network model as a tool for information embedding and extraction and uses the neural network in each stage of embedding and extraction to embed confidential information in multimedia data. In terms of concealment, model steganography has unique advantages compared with its related technologies. The steganography of the model is naturally hidden. The model itself is a complex set of high-dimensional parameters, so a small number of parameter disturbances in the model are difficult to detect. Model steganography is usually achieved by modifying redundant parameters, which will not affect the function of the model. Regarding embedding capacity, model steganography has the potential of supercapacity compared with its related technologies. Model steganography can use parameter redundancy to embed data, and the neural network has a large number of parameters, so it can embed substantial information even if the minimum proportion parameters are modified. In accordance with the different strategies of model steganography, the existing methods can be divided into three categories: model steganography based on training, modification, and backdoor technology. Most of the results of model steganography are training-based model steganography. The main idea of training-based model steganography is to embed confidential information in the process of the training model. In the hidden layer of the model, the sender first selects the weight used to embed the confidential information and then embeds the confidential information into the model under the key function through the training model. In the output layer, the model output is required to be as similar as the confidential information as possible, and the model weight is constantly updated under the guidance of the confidential information. The basic idea of model steganography based on modification is to modify the model parameters to match the confidential information to achieve the purpose of embedding confidential information. Malicious payloads can be embedded without significantly affecting the model performance by replacing malware bytes or mapping model parameters to hide malware in the model. At the sending end, malicious developers choose to modify the location of model parameters to embed malicious software into the model. At the receiving end, they determine the location where the malicious software is embedded in the model parameters, extract the malicious software, check the integrity, and run the malicious software. Model steganography based on backdoor technology uses backdoor technology. Attackers bury backdoors in the model, making the infected model behave normally in general. However, when the backdoor trigger is activated, the output of the model will become the malicious target set by the attacker in advance. This method poisons the target model and can extract additional information from the output of the model. For the analysis method of model steganalysis, on the basis of whether the steganalyzer needs to master the internal details of the neural network model, current model steganalysis algorithms can be classified as white and black box model steganalysis. White box model steganalysis means that the analyst has knowledge and access rights to the internal structure and parameters of the model to detect and analyze the confidential information hidden in the model. Black box model steganalysis treats the target model as a "black box", without accessing its internal structure and weight parameter details, to detect and analyze whether the model contains secret. To review the latest developments and trends, this study analyzes advanced methodologies in model steganography as follows: 1) it introduces the purpose and goal of model steganography, as well as its basic concepts, evaluation indicators, and technology classification. 2) The development status of model steganography is summarized and analyzed. 3) The advantages and disadvantages are compared and evaluated. 4) The development trend of model steganography is explored.

Key words: steganography; model steganography; steganalysis; neural network; information hiding

0 引 言

隐写是保护信息安全的重要手段，将机密隐写到普通媒体中，这种媒体通常称为“载体”，如图像（Song 等，2024）、视频（Meng 等，2024）或文本（Ding 等，2024a）等。隐写的主要目的是实现隐蔽通信，机密信息通过公开信道传递，确保在传输过程中不引起第三方察觉（Wang 等，2022b）。

传统隐写方法具有很大的局限性。如：最低有效位（least significant bit，LSB）替换、离散余弦变换（discrete cosine transform，DCT）和小波分析等方法依赖人工设计特征，易受隐写算法参数变化影响，且对未知数据集的泛化能力较弱，存在低容量与高失真的问题，难以在保持载体质量的同时实现大容量隐写，此外，存在抗分析能力不足、易被统计分析检测以及安全性较低的问题。尽管如此，传统隐写方法仍是轻量化、实时场景的首选。针对上述局限性，很多研究对传统方法进行了改进和优化。Yanuar 等人（2024）将 LSB 与 Josephus 排列结合，能够在保持图像质量的同时，提供更高的有效载荷能力，并提升抗隐写分析能力。Huang 等人（2024b）在 DCT 方法中引入 DCT 残差调制算法，通过量化系数稳定性优化减少误差，结合 RS（Reed-Solomon）编码与单元洗牌分散错误分布，能够提升抗分析能力和鲁棒性。对于小波变换方法，结合 Radon 变换可以增强抗几何攻击能力（El-Den 和 Raslan，2025）。

随着深度学习技术的快速发展和广泛应用，近年来神经网络模型数量增长迅猛。神经网络模型不仅在许多传统任务中表现出优异的性能，如目标跟踪（Wang 和 Song，2025）、图像识别（Shibata 和 Yamauchi，2025）以及自然语言处理（Wu 和 Zhu，2025），而且在隐写术中也表现出强大的性能。神经网络通过学习自动提取特征，能够提升隐写容量、不可检测性和抗隐写分析能力。各种具有较强学习能力的网络模型都可用于隐写，其目的是通过数字媒体传输额外的数据，而不会对载体造成严重的失真，如卷积神经网络（convolutional neural network，CNN）（Liu 等，2022a）、循环神经网络（recurrent neural network，RNN）（Kanimozhi 和 Padmavathi，2025）、生成对抗网络（generative adversarial network，GAN）（Huang 等，2024a）、残差网络（residual network，ResNet）（Liu 等，2022b）以及 U-Net（马宾 等，2024）。自 Transformer 模型提出以来，因其独特的架构设计和强大的性能，在自然语言处理、计算机视觉等领域取得了突破性进展。由于 Transformer 模型出色的长距离建模能力，能够提高图像隐写中局部和全局像素相关性，从而提高含密图像的视觉不可感知性（Dong 等，2024）。Öztürk 等人（2024）在文本隐写中，使用双向编码器表示模型（bidirectional encoder representations from Transformers，BERT）替换掉文本中的特定单词，能够隐藏大量数据而不会扭曲文本含义。大语言模型是基于 Transformer 架构训练的超大规模自然语言处理模型，利用大语言模型可以对文本语义全面控制，尤其是在较长文本的情况下，仍然可以保持连贯性和上下文一致性，有效提高文本隐写的安全性（Li 等，2024c）。

生成式人工智能技术近年来不断突破瓶颈，引起了多领域的革命性变化（严昊 等，2023），也为隐写提供了新的工具和方法。稳定扩散模型（stable diffusion）是最突出的文本到图像生成模型之一，它采用感知压缩进行图像降维，并利用对比语言—图像预训练（contrastive language-image pre-training，CLIP）模型将图像与文本提示对齐。Hu 等人（2024）将机密信息映射到潜空间上，使用稳定扩散模型生成隐写图像，避免修改像素，因此具有不可感知性。神经网络模型用于隐写，需要对高容量与高隐蔽性进行协同优化，同时需要引入对抗性机制以提升抗检测能力，在安全性方面，需要不断开发隐写检测工具，防止模型被恶意滥用。

与图像、音视频和文本等多媒体类似，神经网络模型同样含有大量冗余信息（如不同模型参数可产生相同或相似的模型输出），可为隐藏机密信息提供便利条件。因此，可用神经网络模型为载体，借助神经网络模型传递机密信息，称为神经网络模型隐写（简称模型隐写），如图 1 所示，隐写方在模型中选择嵌入机密信息的位置，使用密钥将机密信息嵌入到模型中传输。而接收方利用共享的密钥，在嵌入机密信息的位置提取出机密信息。

随着神经网络的飞速发展，在神经网络内部直接实现数据隐藏已经越来越普遍。神经网络模型中的数据隐藏可分为模型水印和模型隐写。模型水印将所有者的版权信息嵌入到模型的冗余组件中，以保护知识产权（吴汉舟 等，2023）。首先，模型水印

不能影响模型的原始任务，Wu等人(2021)提出一种数字神经网络水印框架，该框架既可以执行网络的原始任务，又同时将水印嵌入到输出图像中。其次，模型水印需要能够抵抗各种失真，几何攻击会破坏嵌入端和提取端之间的水印同步，是模型水印的重大挑战。为了应对这一挑战，Wang等人(2025)提出使用模板增强提取网络，从扭曲的图像中提取扭曲的模板，再根据扭曲的模板预测的攻击因子，可以实现正确的水印提取和同步，从而抵抗几何攻击。随着大模型时代的崛起，需要对大模型版权进行保护，因此研究人员提出模型确权水印来保护专有大模型(Luo等，2025)。模型水印需要不断增强鲁棒性与抗攻击能力，保证在联邦学习中的可检测性，在保护用户数据隐私的同时嵌入水印，以及实现对于不同模型的通用性。

模型隐写用于无检测的隐蔽通信，近年来神经网络模型隐写技术取得很大进展。在实际中，通过在模型训练过程中嵌入机密信息，或在模型中隐藏执行秘密任务的模型，可以应用在某些场景下，如军事防御或情报机构之间进行秘密通信。在命令分发任务中，指挥官打算向多个军官发送不同的命令，或多个军官向指挥官发送不同的消息，使用模型隐写可以在不被察觉的情况下传递机密信息。与此同时，有些应用场景是不道德甚至非法的，如通过修改模型的参数，恶意开发人员可以将恶意软件嵌入到良性模型中，从而造成模型使用人员的损失；利用神经网络后门技术，对目标模型投毒，可以使得目标模型在不知情的情况下，执行攻击者定义的不同任务，如非法分类或识别、误导决策。研究这些攻击的主要目的是为了提高对潜在威胁的认识，并开发相应的防御机制以保护机器学习系统的安全性和完整性。

近年来，模型隐写逐渐受到关注，已有部分学者开展了模型隐写研究。因此，对模型隐写的发展现状、热点问题和发展趋势进行梳理，可为相关研究人员了解模型隐写的进展提供参考。为此，本文对模型隐写的研究现状进行相对全面的介绍，按照模型隐写的算法原理和特点对目前的成果进行分类和比较，并对模型隐写的发展趋势进行展望，通过这些内容，旨在提供一个关于模型隐写技术的全面视角，展示其在信息安全领域的重要性和潜力。

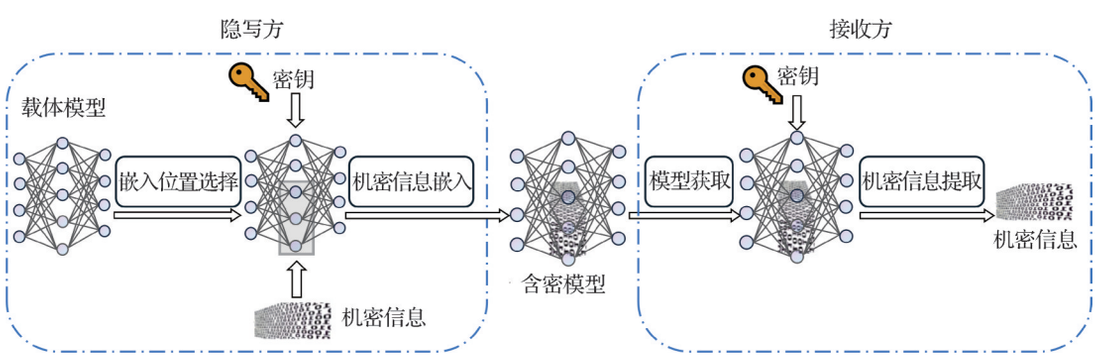


图1　模型隐写一般架构示意图

Fig. 1　General framework for neural network model steganography

1　模型隐写与相关技术的区别

基于上述背景，本节介绍模型隐写与相关技术的区别。模型隐写的相关技术主要为模型水印和基于神经网络模型的多媒体隐写。模型隐写的目的是借助神经网络模型传递机密信息，将机密信息隐藏在神经网络模型中通过公开信道传递。模型隐写技术需要考虑隐蔽性、嵌入容量、鲁棒性、保真度、不可检测性、可恢复性以及计算复杂度等评价指标，各指标的含义如表1所示。

模型隐写与模型水印的区别为：模型水印以神经网络为保护对象，在模型中嵌入数字水印，目的在于保护模型所有者的知识产权(Hassan和Rahma，2024)，可用于图像处理(Quan等，2021)、语音识别(Chen等，2022b)和自然语言处理(He等，2022)等任务。而模型隐写则是将神经网络模型作为隐藏机密信息的载体。在隐蔽性方面，在模型中嵌入的水印信息能被正确提取，且不影响载体模型的正常使用即可，无须刻意隐瞒水印信息的存在性；而模型隐写则侧重于机密信息的隐蔽传输，以不引起第三方察觉为目的，对于隐蔽性的要求远高于模型水印。在

嵌入容量方面，模型水印的容量能容纳水印信息即可，无须追求大容量；而模型隐写在保证隐蔽性的同时，尽量提高嵌入机密信息量，嵌入容量较为重要。在鲁棒性方面，模型水印必须能够抵抗常见的攻击，如模型微调、剪枝等，模型水印对鲁棒性的要求高于模型隐写。

模型隐写与基于神经网络模型的多媒体隐写的区别为：首先，二者负载机密信息的载体不同，基于神经网络模型的多媒体隐写以多媒体数据为载体，如图像（Li 等，2024a）、音视频（Li 等，2023b）和文本（Ding 等，2024b），将神经网络模型作为信息嵌入与提取的工具，在嵌入和提取的各个阶段利用神经网络，从而将机密信息嵌入在多媒体数据中（Dzhanashia 和 Evsutin，2024）。而模型隐写则是以神经网络模型为载体，将机密信息嵌入在模型中。在隐蔽性方面，模型隐写需要确保含密模型与原始模型在统计分布、性能和行为上一致。基于神经网络模型的多媒体隐写需要防止对多媒体数据修改量过大，而引起统计特征异常。在保真度方面，模型隐写需保证精度、推理速度不变，否则会增加暴露风险。而基于神经网络模型的多媒体隐写在保证视觉/听觉质量的情况下可接受轻微失真。在不可检测性方面，基于神经网络模型的多媒体隐写更易实现不可检测性，因有成熟的统计伪装技术；而模型隐写仍处于攻防探索阶段。

在实际应用场景中，模型隐写主要用于隐蔽通信，如通过共享模型传递敏感信息，此外还用于机密信息识别、特定目标识别和特殊文本生成等。模型水印主要用于防止模型盗版，如人工智能（artificial intelligence，AI）服务提供商保护知识产权、追踪模型泄露源等。基于神经网络模型的多媒体隐写主要用于隐蔽数据传输，如通过社交媒体图像传递机密信息。这些技术在实际应用中的优劣势如表 2 所示。

表 1　模型隐写的常见评价指标

Table 1　Common evaluation indicators for neural network model steganography

评价指标	含义
隐蔽性	含密模型不被察觉的能力
嵌入容量	神经网络中含有的机密信息量
鲁棒性	从被攻击的模型中准确提取机密信息的可能性
保真度	含密模型在执行原始任务时的性能损失
不可检测性	含密模型中的机密信息不被现有的隐写分析工具或算法检测出来
可恢复性	接收方从含密模型中准确恢复机密信息的能力
计算复杂度	模型隐写和信息提取时的计算开销

在隐蔽性方面，模型隐写与其相关技术相比有着独特优势。首先，模型隐写具有天然隐蔽性，模型本身是复杂的高维参数集合，因此模型中少量参数扰动难以被察觉。模型隐写通常通过修改冗余参数实现，不会影响模型的功能，其隐蔽性远高于传统多媒体隐写。其次，模型作为合法数字资产能够被广泛地传输和共享，因此隐藏机密信息的行为不易引起怀疑。而模型水印在实际应用中需要抵抗模型微调、剪枝的攻击，可能需要牺牲隐蔽性。基于神经网络模型的多媒体隐写的隐蔽性取决于载体质量，可能因载体的统计特性异常，而被隐写分析工具检测出来。

在嵌入容量方面，模型隐写比其相关技术具有超大容量潜力。模型隐写可以利用参数冗余性嵌入数据，而神经网络的参数量巨大（如 ResNet-50 有 2 500 万参数），因此即使修改极小比例参数也可嵌入大量信息；此外可利用模型结构特性，如权重分布嵌入信息，进一步提升嵌入容量。模型水印的重点在于鲁棒性，嵌入容量通常较低。而基于神经网络

表 2　隐写及相关技术在实际应用中的优劣势比较

Table 2　Advantages and disadvantages of model steganography and related technologies in practical application

技术	优势	不足之处
模型隐写	隐蔽性强、嵌入容量大、载体合法性高	对模型微调敏感、隐藏信息需控制对模型的性能影响、信息提取依赖模型访问权限
模型水印	鲁棒性强、轻量化嵌入、法律认可度高	容量低、隐蔽性较弱
基于神经网络模型的多媒体隐写	兼容传统隐写场景、可结合生成模型提升隐写质量	容量受限于载体分辨率、需平衡隐蔽性与载体质量、易受隐写分析攻击

模型的多媒体隐写受限于载体分辨率，且需平衡载体质量与容量。

2 模型隐写研究现状

目前，模型隐写的研究可大致分为3类：基于训练的模型隐写、基于修改的模型隐写以及基于后门等技术的模型隐写。其中大多数模型隐写的成果为基于训练的模型隐写。

2.1 基于训练的模型隐写

基于训练的模型隐写的主要思想是在训练模型的过程中嵌入机密信息。表3展示了基于训练的模型隐写的定性比较。图2展示了基于训练的模型隐写的过程，发送方在模型的隐藏层中，首先选择用于嵌入机密信息的权重，通过训练模型，在密钥的作用下将机密信息嵌入到模型中。在输出层，要求模型输出与机密信息尽可能相似，在机密信息的引导下不断更新模型权重。

表3 基于训练的模型隐写的定性比较

Table 3 Qualitative comparison of training-based model steganography

方法	改进点	核心思想	侧重点
Wang等人(2021b)	训练策略	用全连接层代替矩阵乘法进行模型参数寻优	嵌入量
Yang等人(2022)	嵌入模式	多个发送者在网络的重叠位置嵌入数据	隐蔽性
Yang等人(2023)	训练策略	用矩阵乘法对卷积层的参数进行编码	嵌入量
Xie和Wang(2024)	提取模式	通过训练含密模型匹配已有的提取器	不可检测性
Hao等人(2025)	嵌入模式	将隐藏层转换为图结构，利用图卷积网络实现数嵌入	鲁棒性
Chen等人(2022a)	嵌入模式	使用一组固定噪声映射的确定性映射来隐藏秘密图像	不可检测性
Guo等人(2021)	嵌入模式	使用密钥对模型参数进行特定的排列	隐蔽性
Guo等人(2022)	训练策略	冻结密钥矩阵，仅优化公开任务的输出层	隐蔽性
Wu等人(2023)	训练策略	串联两个图像输入，经过下采样和上采样层输出	隐蔽性
Li等人(2023a)	训练策略	选择并调整秘密DNN模型中的滤波器子集	不可检测性
Li等人(2024b)	提取模式	生成一组权重来填充网络中的稀疏权重	不可检测性

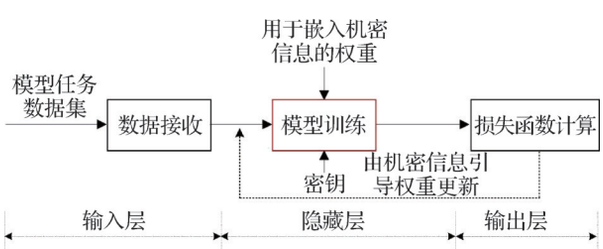


图2 基于训练的模型隐写流程图

Fig. 2 Training-based model steganography

Wang等人(2021b)借助隐写矩阵编码思想，提出一种在胶囊网络(capsule networks，CapsNets)中隐藏机密信息的方法，此方法可以实现多源接收。在发送端，嵌入函数通过矩阵乘法实现，每个接收者对应一个由嵌入密钥生成的嵌入矩阵，用全连接层代替矩阵乘法进行模型参数寻优，嵌入矩阵可以看做是全连接层的参数，在训练过程中将秘密数据嵌入神经网络，可以同时向多个接收者传递不同的秘密数据。在接收端，解码网络设计为全连接层，连接到CapsNets的特定元素(如预测向量)，接收者使用密钥生成解码网络的参数提取秘密数据，而不需要训练，避免了秘密存储和传输解码网络的需求。实验结果表明，所提出的方案能够在不显著降低CapsNets检测精度的情况下，实现对多个接收者的数据隐写，嵌入容量达到6 000 bit，对于命令传递等应用场景是足够的。在公共信道上传输时，含密模型可能受到信道噪声的影响。现有研究中常考虑两类典型噪声：一是服从$N\left(0,\Phi_1^2\right)$分布的高斯噪声，其中Φ_1表示噪声的标准差，即高斯噪声强度；二是服从$U(-\Phi_2,\Phi_2)$分布的均匀噪声，其中Φ_2表示噪声的最大幅值，用于表征均匀噪声的强度。噪声会导致接收端的含密模型参数发生轻微扰动。在对机密信息无误提取的条件下，此方案对上述两种噪声均具有一定的鲁棒性。

Yang等人(2022)提出一种多方发送的模型隐

写方案，多个发送方通过训练模型参数，可向同一接收方传送不同的机密信息，且对原始神经网络的影响很小，具有较好的实用性和安全性。具体而言，多个发送方在神经网络的卷积层权重中嵌入机密信息，每个发送者使用唯一的嵌入密钥，与神经网络的卷积层参数进行点乘运算，并通过sigmoid函数非线性映射混淆嵌入参数。由于密钥的唯一性和随机性，嵌入参数的分布被扰动，从而每个发送方的嵌入数据映射到参数空间的不同子区域，密钥不同，导致参数调整方向各异。在训练过程中，总损失函数的优化会协调多个发送方的嵌入目标，使不同发送方的扰动在参数空间中形成互补或正交的调整方向，因此每个发送方的扰动对其他发送方的数据影响极小，多个发送方信息在接收方能够被准确提取且相互不受干扰。在接收端，接收者使用相应的嵌入密钥和取整操作提取机密信息，而无须额外的解码网络。实验结果表明，对于多个发送者，总嵌入容量与单一发送者一致，每个发送者的嵌入容量随发送者数量增加而减少。在安全性方面，攻击者很难通过网络参数分布来识别机密信息的存在。

在上述方案的基础上，Yang等人（2023）提出对大多数卷积神经网络通用的模型隐写方案，无须针对特定模型调整，此方案能够支持多发送方和多接收方的隐写任务。在模型训练的同时将机密信息嵌入在给定的卷积层中，使用矩阵乘法对卷积层的参数进行编码，嵌入矩阵由密钥生成，将卷积核参数和嵌入矩阵进行矩阵乘法编码，并通过 sigmoid 函数约束输出范围，以便于数据提取，同时可保证原始模型任务的性能。在多源方案中，每个发送者使用独立密钥，生成嵌入矩阵，在网络的重叠位置嵌入数据，多个发送者可以在同一神经网络中嵌入不同的机密信息，接收方需拥有所有密钥来提取全部数据；在多通道方案中，发送者使用多个密钥嵌入不同数据，多个接收者可以从同一神经网络中提取不同的机密信息。实验结果表明，该方法在MNIST（Modified National Institute of Standards and Technology）数据集上使用经典卷积神经网络模型达到了最大5 000 bit的嵌入容量，并且在安全性和鲁棒性方面表现出色。

为了降低提取工具被第三方拦截的风险，Xie和Wang（2024）提出一种提取器匹配的模型隐写方案。接收方无须构建特定的信息提取工具，而是由发送方通过训练模型，匹配接收方已有的神经网络模型，使之可以正确提取机密信息。在训练过程中，将机密信息嵌入到含密模型的权重中。提取密钥确定提取器的参数，这些参数在训练中保持不变，而含密模型的参数则通过优化损失函数进行调整，以确保秘密数据能够被正确提取，同时不影响含密模型的原始任务性能。接收方通过密钥获取提取器的权重，提取器与含密模型的权重首先通过Hadamard乘积和sigmoid函数计算，然后通过舍入操作，再将计算得到的值转换为二进制序列，就得到了机密信息。

该方案在保证提取准确性和网络检测精度的前提下，实现了4 000 bit的容量。在性能上，首先，在安全性方面，第三方即使获得了提取器和含密模型，没有正确的密钥也无法提取有效数据。其次，机密信息是在训练过程中嵌入的，而不是在训练完成后修改参数，确保了网络的原始检测精度不会受到显著影响，从而提高方案的鲁棒性。最后，由于提取器是一个公开可用的普通神经网络，接收方已经拥有，发送方无须传输提取网络，减少了信息泄露的可能性。实验表明，嵌入机密信息前后，模型的参数分布直方图非常相似，难以区分含密模型和普通模型，进一步增强了隐蔽性。然而，不足之处在于此方案的嵌入容量较小，且对于其他类型神经网络的通用性不足。

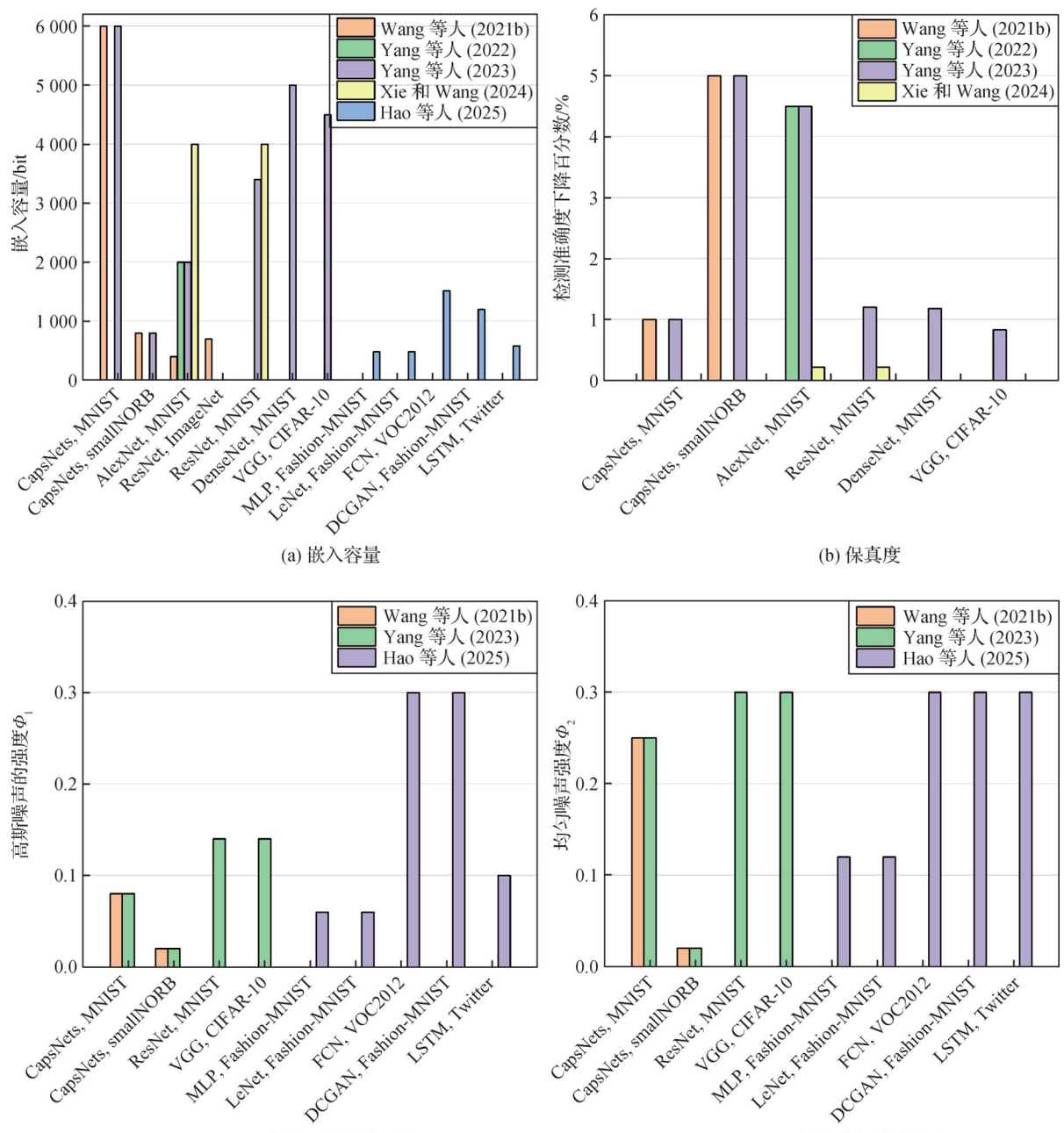
Hao等人（2025）提出一种通用的神经网络模型隐写框架，该方案适用于多种类型的神经网络（如图像分类、语义分割、图像生成和语言生成任务），能够在神经网络的不同层（包括线性层、卷积层和转置卷积层）中嵌入数据。具体而言，通过将神经网络的隐藏层转换为图结构，利用图卷积网络（graph convolutional network，GCN）实现数据嵌入。在嵌入机密信息之前，对GCN进行预训练，使其参数更适合嵌入任务。预训练通过在迭代器中嵌入随机生成的机密信息优化GCN参数，从而减少嵌入过程中对网络权重的过度修改。发送方使用两层GCN进行数据嵌入，将GCN输出限制在[0，1]范围内，再通过最小化机密信息与GCN输出之间的均方误差，确保嵌入机密信息。接收方使用相同的GCN结构和预定义的图连接模式，通过sigmoid函数恢复机密信息。由于GCN的参数可以随机初始化或直接指定，而图结构的连接性由发送方和接收方共同预设，从而避免了传输GCN参数和图连接性，保证了安全性。

该方案由于图结构的引入和预训练，使秘密数

据的嵌入更加均匀，减少了某些权重过度嵌入的情况，从而降低了原始模型和含密模型参数分布之间的差异，提高隐蔽性。其次，通过调整图卷积网络的输出维度或嵌入卷积核的数量，可以实现嵌入容量可变。实验结果表明，当噪声强度较小时，含密模型可以成功提取机密信息；虽然随着噪声强度增加，提取误差逐渐增大，但整体鲁棒性表现良好。

上述提出的各方案，是在模型训练过程中直接嵌入机密信息，通过图3对比各方案在不同数据集、不同模型架构下，在嵌入容量（图3(a)）、保真度（图3(b)）、鲁棒性（图3(c)(d)）上的性能差异。

现有的基于自编码器的隐写方案需要传输较大的解码网络，容易被现有的隐写分析方法识别。为了解决这个问题，Chen等人（2022a）利用模型对原图像集的概率密度函数进行建模，将机密图像隐藏在原图像集概率分布中的某一特定位置，从而将图



(a) 嵌入容量　(b) 保真度

(c) 抵御高斯噪声能力　(d) 抵御均匀噪声能力

图3　基于模型训练嵌入机密信息的各方案性能比较

Fig. 3　Performance comparison of various schemes for embedding confidential information based on model training

((a) embedded capacity; (b) fidelity; (c) ability to resist Gaussian noise; (d) ability to resist uniform noise)

像隐藏在深度概率模型中。具体而言，在分布学习过程中，使用单图像生成对抗网络（single image generative adversarial network，SinGAN）学习载体图像的块分布。SinGAN 由多级生成器和判别器组成，每级生成器通过噪声输入和上采样特征生成图像，而判别器则确保生成图像的块分布与载体图像一致。发送方在 SinGAN 的训练中，将原始重建损失替换为机密图像重建损失，由密钥生成固定噪声映射，将此固定噪声映射与秘密图像相匹配，从而完成图像隐藏过程，训练完成后将含密模型公开。接收方使用嵌入密钥重新生成相同的噪声映射，输入模型中进行一次前向传播，就能提取出机密图像。此方法不需要传输解码网络，而是使用共享的嵌入密钥提取秘密图像，也可以为不同的接收者隐藏多个图像，通过像素混淆来增强安全性。

通过训练模型，不仅可以直接嵌入机密信息，还可以将秘密模型隐藏到一个常规模型中。已经有部分研究关注“以网藏网”的模型隐写，接下来对这些方案进行介绍。

Guo 等人（2021）提出 TrojanNet 框架，该框架通过训练载体模型，利用神经网络的冗余容量同时学习公开任务和秘密任务。具体而言，使用密钥在训练时通过置换操作，打乱网络的权重矩阵，使得秘密任务隐藏在网络内部，可将用于身份识别的网络隐藏在载体模型中。公开任务和秘密任务仅共享参数，不共享特征，得到的含密模型可执行常规的图像分类任务，模型准确率仅略有下降。实验表明，同时训练两个任务后，公开任务的性能几乎不会受到影响。而在没有密钥的情况下，含密模型在秘密任务上的表现与一个随机网络相似，具有隐蔽性。在接收方，使用密钥可从含密模型中提取出隐藏的模型，网络能够有效完成秘密任务，准确率接近专门为秘密任务训练的网络。

Guo 等人（2022）提出一个框架，能够在一个神经网络中同时执行正常任务和隐藏任务。在模型训练过程中，冻结密钥生成的随机矩阵，正常任务使用传统的交叉熵损失训练，而隐藏任务则通过将网络的倒数第 2 层的输出，去乘以密钥生成的随机矩阵激活。网络通过最小化总损失（正常任务损失加隐藏任务损失）优化网络参数，同时保持正常任务的性能不受影响。实验结果表明，当拥有正确密钥时，隐藏任务在不影响原始任务的情况下，可以达到令人满意的性能，保持较高的准确性。当使用错误密钥尝试激活隐藏任务时，隐藏任务的准确性接近 1%，证明隐藏任务无法被检测。

近年来，为了提高隐写性能，越来越多的基于深度神经网络（deep neural network，DNN）的隐写方案相继提出。基于 DNN 的隐写方案通常包含两个主要组成部分：秘密编码（嵌入）网络和秘密解码（恢复）网络。编码网络以载体媒体和秘密数据作为输入，生成隐写媒体；而解码网络则从隐写媒体中提取机密信息。对于基于 DNN 的隐写方案，它显著增加了隐写工具的大小，增加了传输的难度和风险。这引起了人们对如何秘密地、有效地将基于 DNN 的隐写工具传输给发送方和接收方进行秘密通信的关注。

为了解决这一问题，Wu 等人（2023）提出一种方案，将用于图像隐写的网络隐藏在执行图像风格化任务的 DNN 中，该方法不仅隐藏秘密数据的嵌入过程，还隐藏用于嵌入和提取数据的工具。受到 U-Net 启发，在 DNN 模型中通过串联两个图像输入，经过下采样和上采样层，最终生成目标图像。训练后的 DNN 模型可以执行 3 种不同的任务：数据嵌入、数据提取和图像风格化。通过训练 DNN 同时执行这 3 个任务，DNN 的主要任务（图像风格化）掩盖了数据嵌入和提取过程。在数据提取过程中，接收者首先将含密图像与触发信号一起输入到训练好的 DNN 中，然后模型会提取隐藏在含密图像中的机密信息。如果没有触发信号，DNN 无法正确执行提取过程，从而可以避免被检测。这种网络可以在执行正常图像风格化任务的同时，隐藏数据隐写工具，通过特定的模型输入触发模型执行隐写任务。

实验结果表明，通过该方案 DNN 生成的图像在图像风格化、数据嵌入和提取任务中均表现出高质量，虽然随着嵌入量的增加，图像质量略有下降，但仍保持较高的质量，并能够成功提取秘密数据。通过对抗训练，能够显著提高 DNN 对噪声的鲁棒性，即使在噪声干扰的情况下，秘密信息也能有效恢复。此外，通过隐写分析，证明该方法能够有效隐藏数据隐写工具，减少被检测的可能。

Li 等人（2023a）提出一种隐写方案，可以将用于图像隐写解码的 DNN 模型，称为秘密 DNN 模型，嵌入到执行常规机器学习任务的 DNN 模型中。具体而言，首先通过选择和调整对于隐写任务重要但对

常规任务不重要的滤波器作为秘密滤波器，从而保持秘密DNN模型在隐写任务上的功能，然后冻结与秘密任务相关的滤波器，避免其在训练过程中被修改，而仅优化与常规任务相关的滤波器，因此DNN在公开任务上表现良好。通过设计部分参数优化策略，将秘密DNN模型嵌入在神经网络模型中，从而将秘密DNN模型伪装成执行普通机器学习任务的神经网络模型，含密模型可执行常规机器学习任务。在接收端，只有拥有密钥的人员，才能够从含密模型中恢复秘密DNN模型。实验结果表明，所提出的方法在恢复性、保真度、容量和不可检测性方面均表现良好。秘密DNN模型的性能在模型伪装后几乎没有受到影响，含密模型在常规任务上的性能也仅略有下降，难以察觉隐写行为，且参数扩展率小于0.2%，通信负担极小。然而，此方法只能隐藏秘密解码网络。

Li等人(2024b)提出一种名为纯净统一的隐写网络(purified and unified steganographic network，PUSNet)架构，PUSNet能够在一个DNN模型中同时隐藏秘密编码和解码网络，这在实际中更有用。PUSNet有3种工作模式：净化网络模式(PUSNet-P)、秘密编码网络模式(PUSNet-E)和秘密解码网络模式(PUSNet-D)。通过这3种模式，PUSNet可以灵活地在普通机器学习任务和隐写任务之间切换，以实现机密信息的嵌入和恢复。PUSNet的构建问题可形式化为稀疏权重填充问题。净化网络被视为一个稀疏模型，而含密模型是其对应的稠密版本，通过密钥生成的权重填充模型中的稀疏权重，以触发含密模型。

该方案中净化网络在执行普通机器学习任务时，其性能与普通网络几乎一致，二者权重分布距离非常接近，因此可以伪装成一个普通的深度学习模型，增强了隐蔽性。在性能方面，首先，PUSNet通过稀疏权重填充的方式，将含密模型隐藏在净化网络中，这种方法能够灵活调整嵌入容量，通过改变稀疏比例，可以控制含密模型的大小和嵌入容量。其次，在鲁棒性方面，PUSNet能够对抗稀疏权重填充攻击，攻击者通过随机密钥触发秘密网络，生成的图像质量非常差，无法有效恢复秘密数据。净化网络与普通网络在去噪任务上的性能几乎一致，表明PUSNet对噪声具有良好的鲁棒性。此外，PUSNet通过稀疏权重填充的方式隐藏含密模型，使得攻击者即使获取了净化网络，也无法直接提取含密模型，从而提高了PUSNet在模型窃取攻击下的安全性。然而，由PUSNet触发的解码网络进行图像秘密恢复的性能略低于Li等人(2023a)提出的方法。

基于训练的模型隐写方法具有很多优点，首先是具有高隐蔽性，机密信息在训练过程中自然融入模型参数，参数分布与正常模型相似，难以通过统计特征检测。此外鲁棒性较强，对噪声和模型微调具有抗干扰能力。不足之处在于计算成本高，需重新训练模型；其次嵌入容量较小，在容量过大时会导致原始任务精度下降。

2.2 基于修改的模型隐写

基于修改的模型隐写的基本思想是，通过修改模型参数，使其与机密信息匹配，从而达到嵌入机密信息的目的。

Song等人(2017)最先尝试修改模型参数来嵌入秘密消息，提出3种方法：LSB编码、相关值编码和符号编码。LSB编码通过直接修改模型参数的最低有效位编码机密信息；相关值编码通过调整参数值，使得参数与机密信息之间高度相关；符号编码通过调整参数的符号，使参数的符号与机密信息的符号一致。因此恶意开发人员可以使用神经网络在不知不觉中交换消息。然而，此方案的不足之处在于，因模型参数分布异常而容易被检测到，隐蔽性不足。

也有研究人员提出，通过用恶意软件字节替换，或映射模型参数在模型中隐藏恶意软件的方法，可以在不显著影响模型性能的情况下，嵌入恶意有效载荷。嵌入恶意软件的一般流程如图4所示，模型参数为4字节，恶意开发人员在发送端选择修改模型参数的位置，从而将恶意软件嵌入到模型中。在接收端，首先定位到模型参数中嵌入恶意软件的位置，而后提取出恶意软件，并检查完整性，就能够运行恶意软件。表4展示了基于修改的嵌入恶意软件方法的定性比较。

Liu等人(2020)提出一种名为StegoNet的方案，首次将恶意软件隐藏在DNN中。方案中提出4种基于DNN的载荷注入技术：针对未压缩的DNN模型使用LSB替换；针对高度压缩的DNN提出的弹性训练、值映射和符号映射。这些技术利用DNN的结构复杂性、误差容错性和大量参数，将恶意载荷嵌入到DNN模型中，同时保持模型的性能，在现实环境变化下，通过特定物理事件触发StegoNet，提取出恶意

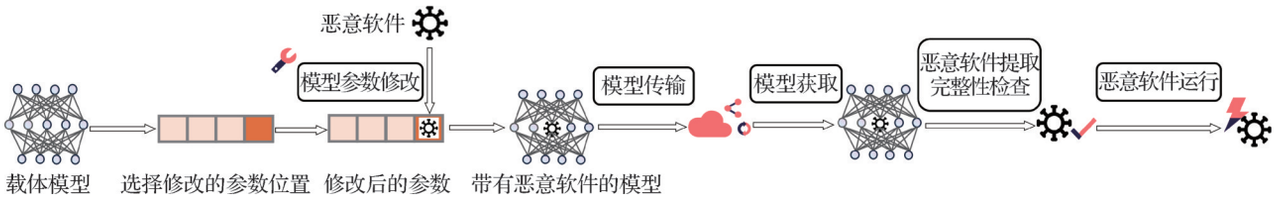


图 4　基于修改的模型隐写（以嵌入恶意软件为例）

Fig. 4　Modifying-based model steganography (embedding malicious software as an example)

表 4　基于修改的嵌入恶意软件方法的定性比较

Table 4　Qualitative comparison of modified embedded malware methods

方法	改进点	核心思想	侧重点
Liu 等人（2020）	参数修改精度	将恶意软件嵌入到参数最低有效位	隐蔽性
Wang 等人（2021a）	参数修改策略	分析参数分布，替换为 3 个字节的恶意软件和 1 个前缀字节	不可检测性
Wang 等人（2022a）	参数修改精度	修改参数后两个字节或后三个字节	嵌入量
Hitaj 等人（2022）	参数修改策略	结合扩频信道编码和奇偶校验纠错技术修改参数	鲁棒性

载荷。实验结果表明，模型的隐蔽性较好，所提出的载荷注入技术，即使在深度压缩的极端条件下，也能保持 DNN 模型参数的恶意载荷完整性。

以上方法存在的问题是：1）嵌入率低，不足以将大型恶意软件嵌入到中小型模型中；2）对模型性能的影响较大，模型的准确性随着恶意软件大小的增加而显著下降，特别是对于小型模型。这些问题阻碍了其在现实的有效使用。

针对上述问题，Wang 等人（2021a）提出一种名为“快速替换”（fast substitution）的方法传递恶意软件。这种方法通过修改整个神经元嵌入恶意软件。由于神经网络层中存在冗余神经元，改变某些神经元对网络性能的影响很小，同时能够保持模型结构不变。具体而言，快速替换方法通过分析参数分布，将恶意软件字节转换为合理的浮点数，攻击者根据参数值，将参数替换为 3 个字节的恶意软件和 1 个前缀字节。发送方使用每个神经元中的连接权重存储转换后的恶意软件字节，而使用偏差存储恶意软件的长度和哈希值。提取是一个反向嵌入的过程，接收方需要提取给定层神经元的参数，将参数转换为浮点数，对恶意软件进行组装，再通过将提取的恶意软件的哈希值与偏差中记录的哈希值进行比较，从而验证提取过程。实验结果表明，快速替换比 StegoNet 具有更高的嵌入容量，并且可以躲过普通杀毒引擎的安全扫描。

随后，Wang 等人（2022a）进一步提出另外两种嵌入方法：最高有效字节（most significant byte，MSB）的保留和半替换。MSB 保留方法是保持参数第 1 个字节不变，在最后 3 个字节中嵌入恶意软件；MSB 半替换方法是保持参数前两个字节不变，修改其余两个字节。实验表明，通过在线杀毒软件扫描和隐写分析，都不能有效检测到嵌入的恶意软件，模型熵检查也没有发生明显变化，证明了恶意模型的有效性和隐蔽性。与 StegoNet 相比，Wang 等人（2021a，2022a）提出的 3 种新方法能够实现更高的嵌入率，能够将几乎是模型体积一半的恶意软件嵌入到模型中，而不影响模型性能。此外，该方案也无需重新训练或索引排列，这在实际应用中更为方便。文中提出一种定量方法来评估和比较现有的嵌入方法，该方法结合了嵌入率、模型性能影响和嵌入工作量。评估结果表明，半替换方法在所有评估指标上都优于 StegoNet。

然而，Wang 等人（2021a，2022a）提出的 3 种恶意软件嵌入方法的鲁棒性较差，要求嵌入恶意软件的模型不能被修改。如果神经元的参数通过微调、剪枝、模型压缩或其他操作而改变，就会破坏恶意软件结构，从而使恶意软件无法正常恢复。

为了缓解上述问题，Hitaj 等人（2022）提出一种新的有效载荷嵌入技术 MaleficentNet，利用码分多址扩频信道编码和低密度奇偶校验纠错技术，将恶意软件嵌入到不同的 DNN 体系结构中。MaleficentNet 通过将恶意负载分成多个块，并利用码分多址扩

频信道编码，将每个块的负载嵌入到DNN的权重参数中。在不显著改变DNN权重参数的情况下，MaleficentNet可以嵌入恶意负载，由于对恶意负载进行了纠错编码，即使DNN参数被修改，也能保持负载的完整性。实验结果表明，MaleficentNet能够在不显著降低模型性能的情况下，成功嵌入不同大小的恶意负载。此外，嵌入恶意负载的模型能够成功绕过最先进的恶意软件检测引擎，同时对微调和参数修剪等去除技术具有鲁棒性。

基于修改的模型隐写方法的优点如下，首先实现较简单，无须重新训练模型，而是直接修改参数；此外灵活性强，可根据需求选择参数修改位置；资源消耗低，适合资源受限场景。而不足之处在于隐蔽性差，参数分布易异常，容易被隐写分析器检测出来；此外鲁棒性差，对模型剪枝、压缩或微调敏感，需要依赖编码技术增强抗修改能力。

2.3 基于后门等技术的模型隐写

也有学者研究利用后门等技术隐藏信息的模型隐写方法。“后门攻击”是一种针对机器学习模型的攻击方式，攻击者会在模型中埋藏后门，使得被感染的模型（infected model）在一般情况下表现正常。但当后门触发器被激活时，模型的输出将变为攻击者预先设置的恶意目标。由于模型在后门未被触发之前表现正常，因此这种恶意的攻击行为很难被发现。

Salem等人（2021）利用神经网络后门技术，对目标模型进行投毒，提出一种可在模型的输出中提取额外信息的方案，而被投毒的模型可正常执行常规任务，如图5所示。这是第一个针对机器学习模型的模型劫持攻击，在这种攻击中，攻击者通过在训练数据中注入特定的数据，使得目标模型在不知情的情况下，执行攻击者定义的不同任务。首先，在劫持任务和原始任务之间建立标签映射，将劫持任务的标签映射到原始任务标签。然后通过伪装器（camouflager）生成视觉相似、语义保留的伪装数据集，结合视觉和语义的双损失优化实现隐蔽劫持。伪装器是基于编码器—解码器的生成模型，该模型包含两个编码器，一个用于编码目标模型的目标数据集（视觉上类似原始数据集），另一个用于编码攻击者的劫持数据集。这两个编码器的输出串接并输入同一个解码器中，生成的伪装数据集，在视觉上与目标数据集相似，但在语义上与劫持数据集相似。最后，将伪装数据集注入目标模型的原始数据集，训练被劫持的模型。

在执行攻击时，先伪装劫持样本，再输入被劫持模型，将预测标签反向映射回劫持任务标签，获得劫持任务结果。在变色龙攻击（chameleon attack）中出现的相似数据集场景的混淆问题，反变色龙攻击（adverse chameleon attack）可通过对抗语义损失来解决。实验结果表明，这种方案在保持目标模型在原始任务上的性能的同时，能够成功执行劫持任务，并在多个数据集上验证了攻击的有效性，同时也暴露出了联邦学习等开放训练范式的安全隐患。

基于后门等技术的模型隐写方法具有极强隐蔽性，模型在正常任务中表现无异，仅通过特定触发器激活。不足是这种方法依赖触发器设计，关键在于触发器的隐蔽性和多样性，设计不当则容易被防御机制识别。

2.4 模型隐写分析

相对地，模型隐写分析旨在通过分析神经网络模型的输入、输出，以及模型参数，判断可疑模型中

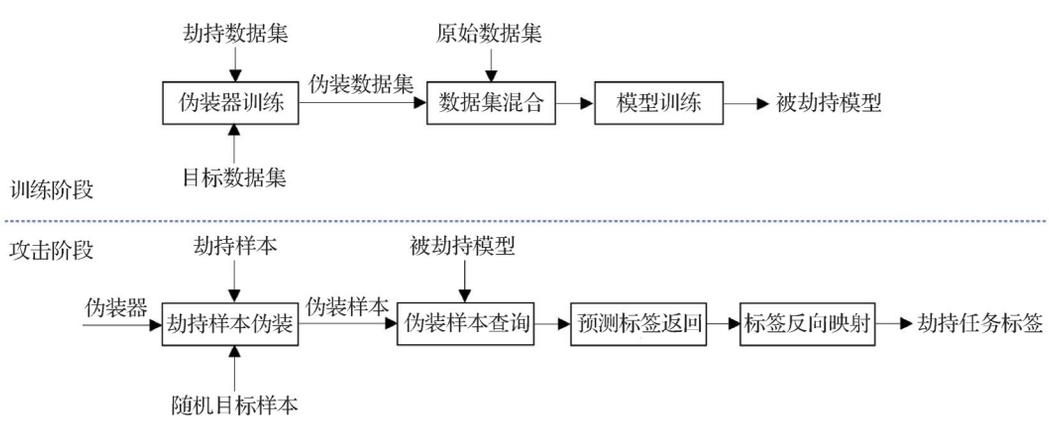


图5 基于后门技术的模型隐写流程图

Fig. 5 Model steganography based on backdoor technology

是否含有机密信息。由于目前对模型隐写的研究较少，模型隐写分析研究也刚刚开始，根据隐写分析者是否需要掌握神经网络模型的内部细节，可将当前模型隐写分析算法归类为白盒模型隐写分析和黑盒模型隐写分析。

2.4.1　白盒模型隐写分析

白盒模型隐写分析是指分析者对模型的内部结构和参数有了解和访问权限，以检测和分析隐藏在模型中的机密信息。白盒模型隐写分析的一般性框架如图6所示。

AI模型共享和下载是一种新的攻击媒介。许多平台提供了各种免费的最先进的预训练模型，可供任何人下载，但是可能被黑客用来进行恶意攻击，例如在2.2节中提及的在神经网络中嵌入恶意软件。

为了检测恶意开发者可能在模型参数中隐藏恶意软件或其他有害信息，保护共享模型的完整性，维护用户信任，并保持AI社区开放合作的良好氛围，尹奕等人(2022)首次提出对神经网络参数隐写的一种隐写分析方法，具体而言，该方案基于由隐写行为造成的偏差，包括模型参数位平面随机性偏差与参数分布的偏差，从正常模型与含密模型中提取特征，再建立分类器以实现隐写分析。针对3种已知隐写方法，分别是最低有效位编码、相关值编码和符号编码，设计了不同的特征提取策略：由于最低有效位编码会修改低位比特，因此通过分析参数低位比特的随机性来提取特征；而相关值编码会约束参数值范围，因此通过分析参数数值分布的统计矩来提取特征；符号编码会强制参数符号，因此可通过分析参数符号分布的统计矩提取特征。针对未知隐写方法，可通过分类器融合实现综合检测，采取特征联合和投票机制的融合策略，同时提取3类隐写方法的特征，对所有子分类器的预测结果通过投票融合，仅当所有子分类器都判定为正常时，才最终判定为正常模型。对于未知隐写方法的场景，该方案能够为AI模型安全提供多维度保障。然而，如果通过降低神经网络中的嵌入率减少差异，就能逃避这种方法的检测。

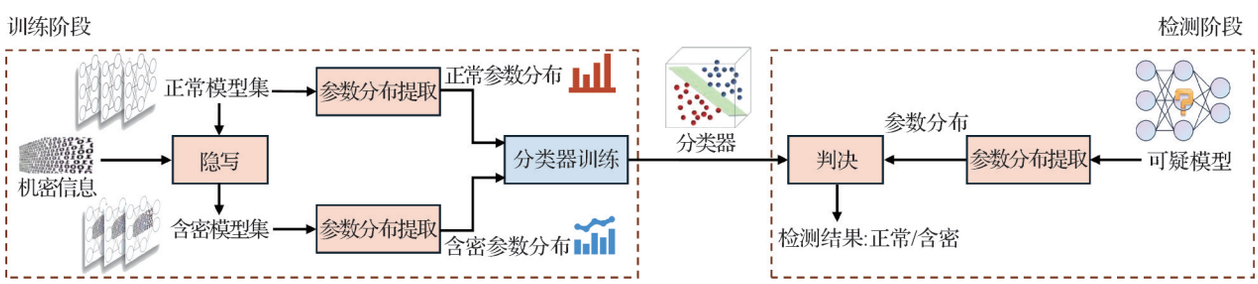


图6　白盒模型隐写分析一般框架

Fig. 6　General framework for white-box analysis of model steganography

为了解决尹奕等人(2022)提出的神经网络隐写分析方法在小嵌入率下检测性能不佳的问题，Zhao等人(2023)提出一种基于校准的模型隐写分析方法，该方案主要针对相关值编码和符号编码隐写。具体而言，首先进行校准，对于待检测模型进行无隐式约束微调，得到参考模型，以消除隐写导致的参数分布偏移。再从二者的各层参数分布中提取四阶统计矩(期望、方差、偏度和峰度)，从而覆盖参数分布的多维度信息。接着进行特征融合，即将原始特征与校准特征的差值拼接，因此能够保留分布变化的敏感信息，增强对由于隐写引入的微小偏差的捕捉能力，最后训练逻辑回归分类器，以检测模型是否含密。实验结果表明，Zhao等人(2023)方法在小嵌入率下检测隐写神经网络模型的性能显著优于尹奕等人(2022)方法。

Gilkarov和Dubin(2024)提出一个在深度学习模型上检测LSB隐写攻击的隐写分析工作，提出3种特征提取方法，包括使用重建损失、反向传播梯度和模型权重作为特征，用于在监督和无监督学习中识别LSB攻击。第1种方法利用重建损失作为特征，是通过自编码器学习良性模型权重的分布特征，计算输入模型的重建误差。由于被攻击模型的权重分布偏离正常，会导致重建损失升高；第2种方法利用反向传播过程中梯度作为统计特征，输入零值数据，以比较良性模型与被攻击模型的梯度差异，被攻击模型的梯度分布会因权重篡改而异常；第3种方

法直接以模型权重的高维向量作为特征，捕捉全局篡改痕迹，这种方法适用于攻击区域覆盖大部分权重的情况，然而对计算资源要求较高。对于监督学习，是从良性模型和被攻击模型中提取特征，构建带标签的训练集，再训练分类器，以检测被攻击模型；对于无监督学习，是将基于良性数据的特征均值设定为阈值，若超出阈值则判定为恶意。该方案构建了一个专门用于隐写分析的学习数据集的新框架。

该方案中隐写分析的具体做法是，首先构建被攻击的模型数据集，对原始模型参数的不同位置进行攻击，即在模型的权重的不同位置中嵌入恶意字符串。然后对原始模型和被攻击的模型提取特征，创建一个数据集，其中每个模型对应一个特征值。这个数据集用于训练和测试隐写分析模型，以区分良性模型和受到攻击的模型。实验结果表明，权重被攻击的位数越高，检测难度越低。此方案对于攻击者针对模型参数中较低位的攻击检测效果较差，并且需要大量的训练数据，计算量大。

2.4.2 黑盒模型隐写分析

黑盒模型隐写分析是将目标模型视为“黑盒”，不需要访问其内部结构和权重参数细节，以检测和分析模型是否含密。Tang等人(2023)提出一种黑盒模型隐写分析方法，是专门为执行图像分类任务的神经网络设计的对称隐写分析方案。图7给出了该方法对应的一般性框架。该方案构建了“隐写特征提取+集成分类器训练”的神经网络隐写特征盲检测框架，通过设计一种基于直方图分布的方法，找到最优的决策阈值，以确定模型是否包含机密信息。

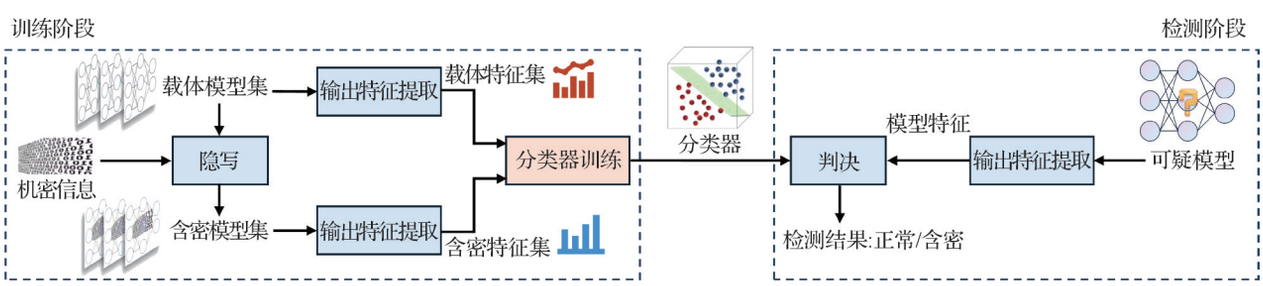


图7 黑盒模型隐写分析一般框架

Fig. 7 General framework for black-box analysis of model steganography

该方案通过分析模型的输入—输出行为提取有效特征，用于检测模型是否含密。其核心步骤如下：1)构建数据集。首先，创建一个包含普通模型和含密模型的数据集。普通模型是为图像分类任务训练的标准模型，而含密模型则是在训练过程中嵌入了机密信息的模型。为了确保数据集的多样性，使用了不同类型的模型架构，并在多个图像数据集上进行训练。同时，含密模型的嵌入容量范围设定为几百比特到几千比特，以涵盖不同的信息嵌入情况。2)提取特征。将模型的输出预测向量作为特征向量，这些向量是在模型完成原始分类任务时产生的，因此无须访问模型的内部结构或参数。为了提高后续分类器的训练效率和检测性能，对提取出的特征向量进行归一化处理。3)训练分类器。选择支持向量机(support vector machine，SVM)和集成分类器作为分类器。通过交叉验证选择合适的核函数(如高斯核、多项式核等)，以平衡模型复杂度和训练数据的拟合程度。最后，确定最优决策阈值。通过分析普通模型和含密模型输出预测标签的分布，找到两者的差异并确定分类边界。具体做法是：将两类模型的预测标签序列分别分割成多个连续段；为每个段生成直方图并拟合其概率密度曲线；通过寻找密度曲线中两个峰值之间的交点，确定该段的最优分类阈值。将所有段的最优阈值组合成一个综合阈值向量，作为检测的决策依据。

对于待检测的模型，首先输入一组测试图像，收集模型输出的预测向量作为特征向量。然后，将这些特征向量进行相同的归一化处理，并输入到训练好的SVM和集成分类器中，得到预测的标签向量。接着，将预测的标签向量分割成多个段，段数需与综合阈值向量中的阈值数量一致，且每段长度相同。统计每个段中标签为“1”(表示预测为含密)的数量。最后，将每个段统计得到的“1”的数量，与该段在综合阈值向量中对应的阈值进行比较。根据比较结

果，采用多数投票机制，最终判定该模型是否含有机密信息。

3 结　语

从模型隐写的主要指标来看，当前研究人员正不断研究更有效的算法，以提高机密信息的安全性和隐蔽性，同时努力提高模型在一些攻击（如模型剪枝和压缩）下的鲁棒性，以确保机密信息的完整性和可提取性，并在保证隐蔽性的情况下提升嵌入量。

神经网络模型隐写是一个新兴的研究方向，该方向刚刚起步，就受到了全世界的关注。然而研究成果还远远不够，亟待学术界开展深入研究，在理论基础、技术手段和实现方法等方面取得新突破。目前亟待发展之处包括但不限于以下 3 点：1）目前提出的模型隐写方法主要针对参数量相对较小的神经网络模型，大模型的参数规模与复杂度远高于普通模型，通过重新训练大模型隐藏机密信息等模型隐写策略，在大模型中执行极为困难。在未来的研究中，发展适用于大模型的隐写方法意义重大。2）当前的模型隐写方案对模型原始任务性能有少许影响，且未讨论抗模型隐写分析检测的性能，隐写嵌入容量也有待提升，因此，大容量高隐蔽的模型隐写是一个值得研究的方向。3）隐写对隐蔽性要求较高，现有模型隐写方案并未关注除机密信息本身以外的隐蔽性。在真实场景中，嵌入操作的实施、含密模型的传送，以及机密信息的提取等环节，均有导致隐写失败的风险，因此，全方位安全的模型隐写也是值得期待的研究。

参考文献（References）

Chen H Y, Song L Q, Qian Z X, Zhang X P and Ma K D. 2022a. Hiding images in deep probabilistic models//Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS 2022). New Orleans, USA: Neural Information Processing Systems Foundation: 36776-36788

Chen H Z, Zhang W M, Liu K L, Chen K J, Fang H and Yu N H. 2022b. Speech pattern based black-box model watermarking for automatic speech recognition//Proceedings of 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapore, Singapore: IEEE: 3059-3063 [DOI: 10.1109/ICASSP43922.2022.9747044]

Ding C H, Fu Z J, Yang Z L, Yu Q, Li D Q and Huang Y F. 2024a. Context-aware linguistic steganography model based on neural machine translation. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 32: 868-878 [DOI: 10.1109/TASLP.2023.3340601]

Ding C H, Fu Z J, Yu Q, Wang F and Chen X Y. 2024b. Joint linguistic steganography with Bert masked language model and graph attention network. IEEE Transactions on Cognitive and Developmental Systems, 16(2): 772-781 [DOI: 10.1109/TCDS.2023.3296413]

Dong Y Y, Wei P, Wang R X, Song B B, Wei T C and Zhou W. 2024. Hiding image with inception transformer. IET Image Processing, 18(13): 3961-3975 [DOI: 10.1049/ipr2.13225]

Dzhanashia K and Evsutin O. 2024. Neural networks-based data hiding in digital images: overview. Neurocomputing, 581: #127499 [DOI: 10.1016/j.neucom.2024.127499]

El-Den B M and Raslan W. 2025. A reversible and robust hybrid image steganography framework using radon transform and integer lifting wavelet transform. Scientific Reports, 15(1): #15687 [DOI: 10.1038/s41598-025-98539-2]

Gilkarov D and Dubin R. 2024. Steganalysis of AI models LSB attacks. IEEE Transactions on Information Forensics and Security, 19: 4767-4779 [DOI: 10.1109/TIFS.2024.3383770]

Guo C, Wu R H and Weinberger K Q. 2021. On hiding neural networks inside neural networks [EB/OL]. [2024-11-24]. https://arxiv.org/pdf/2002.10078.pdf

Guo Y S, Qian Z X and Zhang X P. 2022. Hiding function with neural networks//Proceedings of the 24th IEEE International Workshop on Multimedia Signal Processing (MMSP). Shanghai, China: IEEE: 1-5 [DOI: 10.1109/MMSP55362.2022.9949163]

Hao Y L, Wang Z C, Cao J M and Zhang X P. 2025. General steganography for neural network models based on graph convolutional network. IEEE Internet of Things Journal, 12(9): 12512-12526 [DOI: 10.1109/JIOT.2024.3520994]

Hassan Y A and Rahma A M S. 2024. Improving video watermarking through galois field $GF(2^4)$ multiplication tables with diverse irreducible polynomials and adaptive techniques. CMC-Computers, Materials and Continua, 78(1): 1423-1442 [DOI: 10.32604/cmc.2023.046149]

He X L, Xu Q K, Lyu L J, Wu F Z and Wang C G. 2022. Protecting intellectual property of language generation APIs with lexical watermark//Proceedings of the 36th AAAI Conference on Artificial Intelligence (AAAI 2022). Palo Alto, USA: AAAI Press: 10758-10766 [DOI: 10.1609/aaai.v36i10.21321]

Hitaj D, Pagnotta G, Hitaj B, Mancini L V and Perez-Cruz F. 2022. MaleficNet: hiding malware into deep neural networks using spread-spectrum channel coding//Proceedings of the 27th European Symposium on Research in Computer Security (ESORICS). Copenhagen, Denmark: Springer: 425-444 [DOI: 10.1007/978-3-031-

17143-7_21]

Hu X X, Li S, Ying Q C, Peng W L, Zhang X P and Qian Z X. 2024. Establishing robust generative image steganography via popular stable diffusion. IEEE Transactions on Information Forensics and Security, 19: 8094-8108 [DOI: 10.1109/TIFS.2024.3444311]

Huang D X, Luo W Q, Liu M L, Tang W X and Huang J W. 2024a. Steganography embedding cost learning with generative multi-adversarial network. IEEE Transactions on Information Forensics and Security, 19: 15-29 [DOI: 10.1109/TIFS.2023.3318939]

Huang Y K, Liu Z X, Wu Q W and Liu X L. 2024b. Robust image steganography against JPEG compression based on DCT residual modulation. Signal Processing, 219: #109431 [DOI: 10.1016/j.sigpro.2024.109431]

Kanimozhi R and Padmavathi V. 2025. Robust and secure image steganography with recurrent neural network and fuzzy logic integration. Scientific Reports, 15(1): #13122 [DOI: 10.1038/s41598-025-97795-6]

Li F Y, Sheng Y, Zhang X P and Qin C. 2024a. iSCMIS: spatial-channel attention based deep invertible network for multi-image steganography. IEEE Transactions on Multimedia, 26: 3137-3152 [DOI: 10.1109/TMM.2023.3307970]

Li G B, Li S, Li M L, Zhang X P and Qian Z X. 2023a. Steganography of steganographic networks//Proceedings of the 39th AAAI Conference on Artificial Intelligence (AAAI 2023). Philadelphia, Pennsylvania, USA: Association for the Advancement of Artificial Intelligence (AAAI): 5178-5186 [DOI: 10.1609/aaai.v37i4.25647]

Li G B, Li S, Luo Z C, Qian Z X and Zhang X P. 2024b. Purified and unified steganographic network//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE Computer Society: 27559-27568 [DOI: 10.1109/CVPR52733.2024.02603]

Li Y H, Zhang R, Liu J Y and Lei Q. 2024c. A semantic controllable long text steganography framework based on LLM prompt engineering and knowledge Graph. IEEE Signal Processing Letters, 31: 2610-2614 [DOI: 10.1109/LSP.2024.3456636]

Li Z H, Jiang X H, Dong Y, Meng L J and Sun T F. 2023b. An anti-steganalysis HEVC video steganography with high performance based on CNN and PU partition modes. IEEE Transactions on Dependable and Secure Computing, 20(1): 606-619 [DOI: 10.1109/TDSC.2022.3140899]

Liu J D, Li Z H, Jiang X H and Zhang Z Z. 2022a. A high-performance CNN-applied HEVC steganography based on diamond-coded PU partition modes. IEEE Transactions on Multimedia, 24: 2084-2097 [DOI: 10.1109/TMM.2021.3075858]

Liu L S, Meng L Z, Wang X L and Peng Y J. 2022b. An image steganography scheme based on ResNet. Multimedia Tools and Applications, 81(27): 39803-39820 [DOI: 10.1007/s11042-022-13206-2]

Liu T, Liu Z H, Liu Q, Wen W J, Xu W Y and Li M. 2020. STegoNeT: turn deep neural network into a stegomalware//Proceedings of the 36th Annual Computer Security Applications Conference (ACSAC). Austin, USA: Association for Computing Machinery (ACM): 928-938 [DOI: 10.1145/3427228.3427268]

Luo H X, Li L and Li J C. 2025. Digital watermarking technology for AI-generated images: a survey. Mathematics, 13(4): #651 [DOI: 10.3390/math13040651]

Ma B, Li K, Xu J, Wang C P, Li J and Zhang L W. 2024. High-security image steganography with the combination of multiple competition and channel attention. Journal of Image and Graphics, 29(2): 355-368 (马宾, 李坤, 徐健, 王春鹏, 李健, 张立伟. 2024. 联合多重对抗与通道注意力的高安全性图像隐写. 中国图象图形学报, 29(2): 355-368) [DOI: 10.11834/jig.230134]

Meng L J, Jiang X H, Sun T F, Zhao Z Y and Xu Q. 2024. A robust coverless video steganography based on the similarity of inter-frames. IEEE Transactions on Multimedia, 26: 5996-6011 [DOI: 10.1109/TMM.2023.3344357]

Öztürk E, Mesut A Ş and Fıdan Ö A. 2024. A character based steganography using masked language modeling. IEEE Access, 12: 14248-14259 [DOI: 10.1109/ACCESS.2024.3354710]

Quan Y H, Teng H, Chen Y X and Ji H. 2021. Watermarking deep neural networks in image processing. IEEE Transactions on Neural Networks and Learning Systems, 32(5): 1852-1865 [DOI: 10.1109/TNNLS.2020.2991378]

Salem A, Backes M and Zhang Y. 2021. Get a model! model hijacking attack against machine learning models//Network and Distributed System Security Symposium. San Diego, USA: [s.n.]

Shibata R and Yamauchi Y. 2025. End-to-end learning framework incorporating image reconstruction and recognition models. IEEE Access, 13: 73355-73361 [DOI: 10.1109/ACCESS.2025.3563476]

Song B B, Wei P, Wu S X, Lin Y and Zhou W. 2024. A survey on deep-learning-based image steganography. Expert Systems with Applications, 254: #124390 [DOI: 10.1016/j.eswa.2024.124390]

Song C Z, Ristenpart T and Shmatikov V. 2017. Machine learning models that remember too much//Proceedings of the 24th ACM SIGSAC Conference on Computer and Communications Security (ACM CCS). Dallas, USA: Association for Computing Machinery (ACM): 587-601 [DOI: 10.1145/3133956.3134077]

Tang X, Wang Z C and Zhang X P. 2023. Steganalysis of neural networks based on symmetric histogram distribution. Symmetry (Basel), 15(5): #1079 [DOI: 10.3390/sym15051079]

Wang H and Song L P. 2025. Extended target tracking using neural network and Gaussian process. Electronics Letters, 61(1): #e70151 [DOI: 10.1049/ell2.70151]

Wang K, Wu S W, Yin X L, Lu W, Luo X Y and Yang R. 2025. Robust image watermarking with synchronization using template enhanced-extracted network. IEEE Transactions on Circuits and Systems for Video Technology, 35(2): 1602-1614 [DOI: 10.1109/TCSVT.2024.3474029]

Wang Z, Liu C G and Cui X. 2021a. EvilModel: hiding malware inside of neural network models//Proceedings of the 26th IEEE Symposium on Computers and Communications (ISCC 2021). Athens, Greece: IEEE: 1-7 [DOI: 10.1109/ISCC53001.2021.9631425]

Wang Z, Liu C G, Cui X, Yin J and Wang X T. 2022a. EvilModel 2.0: bringing neural network models into malware attacks. Computers and Security, 120: #102807 [DOI: 10.1016/j.cose.2022.102807]

Wang Z C, Feng G R, Wu H Z and Zhang X P. 2021b. Data hiding in neural networks for multiple receivers. IEEE Computational Intelligence Magazine, 16(4): 70-84 [DOI: 10.1109/MCI.2021.3108305]

Wang Z C, Feng G R and Zhang X P. 2022b. Repeatable data hiding: towards the reusability of digital images. IEEE Transactions on Circuits and Systems for Video Technology, 32(1): 135-146 [DOI: 10.1109/TCSVT.2021.3057599]

Wu D Q and Zhu C. 2025. Interactive memory networks based on syntactic dependencies for aspect-level sentiment classification. The Journal of Supercomputing, 81(1): #189 [DOI: 10.1007/s11227-024-06594-9]

Wu H Z, Li C, Liu G and Zhang X P. 2023. Hiding data hiding. Pattern Recognition Letters, 165: 122-127 [DOI: 10.1016/j.patrec.2022.12.008]

Wu H Z, Liu G, Yao Y W and Zhang X P. 2021. Watermarking neural networks with watermarked images. IEEE Transactions on Circuits and Systems for Video Technology, 31(7): 2591-2601 [DOI: 10.1109/TCSVT.2020.3030671]

Wu H Z, Zhang J, Li Y, Yin Z X, Zhang X P, Tian H, et al. 2023b. Overview of artificial intelligence model watermarking. Journal of Image and Graphics, 28(6): 1792-1810 (吴汉舟, 张杰, 李越, 殷赵霞, 张新鹏, 田晖, 等. 2023. 人工智能模型水印研究进展. 中国图象图形学报, 28(6): 1792-1810) [DOI: 10.11834/jig.230010]

Xie Y F and Wang Z C. 2024. Neural network steganography using extractor matching//Proceedings of the 22nd International Workshop on Digital-Forensics and Watermarking (IWDW 2023). Jinan, China: Springer: 169-179 [DOI: 10.1007/978-981-97-2585-4_12]

Yan H, Liu Y L, Jin L W and Bai X. 2023. The development, application, and future of LLM similar to ChatGPT. Journal of Image and Graphics, 28(9): 2749-2762 (严昊, 刘禹良, 金连文, 白翔. 2023. 类 ChatGPT 大模型发展、应用和前景. 中国图象图形学报, 28(9): 2749-2762) [DOI: 10.11834/jig.230536]

Yang Z Y, Wang Z C and Zhang X P. 2023. A general steganographic framework for neural network models. Information Sciences, 643: #119250 [DOI: 10.1016/j.ins.2023.119250]

Yang Z Y, Wang Z C, Zhang X P and Tang Z J. 2022. Multi-source data hiding in neural networks//Proceedings of the 24th IEEE International Workshop on Multimedia Signal Processing (MMSP 2022). Shanghai, China: IEEE: 1-6 [DOI: 10.1109/MMSP55362.2022.9948867]

Yanuar M R, Suryadi M T, Apriono C and Syawaludin M F. 2024. Image-to-image steganography with josephus permutation and least significant bit (LSB) 3-3-2 embedding. Applied Sciences (Basel), 14(16): #7119 [DOI: 10.3390/app14167119]

Yin Y, Zhang W M, Yu N H and Chen K J. 2022. Steganalysis of neural networks based on parameter statistical bias. Journal of University of Science and Technology of China, 52(1): 1-1-1-12 (尹奕, 张卫明, 俞能海, 陈可江. 2022. 基于参数特征偏移的神经网络隐写检测方法. 中国科学技术大学学报, 52(1): 1-1-1-12) [DOI: 10.52396/JUSTC-2021-0197]

Zhao N, Chen K J, Qin C, Yin Y, Zhang W M and Yu N H. 2023. Calibration-based steganalysis for neural network steganography//Proceedings of the 11th ACM Workshop on Information Hiding and Multimedia Security (IH and MMSec). Chicago, USA: Association for Computing Machinery: 91-96 [DOI: 10.1145/3577163.3595100]

作者简介

龙玲慧，女，硕士研究生，主要研究方向为隐写和隐写分析。
E-mail: lhloong@shu.edu.cn

王子驰，通信作者，男，副研究员，主要研究方向为隐写、隐写分析和人工智能安全。E-mail: wangzichi@shu.edu.cn

张新鹏，男，教授，主要研究方向为多媒体信息安全。
E-mail: xzhang@shu.edu.cn