

基于黑盒水印的 NLP 神经网络版权保护

代龙, 张静, 樊雪峰, 周晓谊

(海南大学网络空间安全学院, 海南 海口 570100)

摘要: 随着自然语言处理 (NLP, natural language processing) 技术的快速发展, 语言模型在文本分类和情感分析中的应用不断增加。然而, 语言模型容易遭到盗版再分发, 对模型所有者的知识产权造成严重威胁。因此, 研究者着手设计保护机制来识别语言模型的版权信息。现有的适用于文本分类任务的语言模型水印无法与所有者身份相关联, 且鲁棒性不足以及无法再生成触发集。为了解决这些问题, 提出一种新的适用于文本分类任务模型的黑盒水印方案, 可以远程快速验证模型所有权。将模型所有者的版权消息和密钥通过密钥相关的哈希运算消息认证码 (HMAC, hash-based message authentication code) 得到版权消息摘要, 由 HMAC 得到的消息摘要可以防止被伪造, 具有很强的安全性。从原始训练集各个类别中随机挑选一定的文本数据, 将摘要与文本数据结合构建触发集, 并在训练过程中对语言模型嵌入水印。为了评估水印的性能, 在 IMDB 电影评论、CNEWS 中文新闻文本分类数据集上对 3 种常见的语言模型嵌入水印。实验结果表明, 在不影响原始模型测试精度的情况下, 所提出的水印验证方案的准确率可以达到 100%。即使在模型微调 and 剪枝等常见攻击下, 也能表现出较强的鲁棒性, 并且具有抗伪造攻击的能力。同时, 水印的嵌入不会影响模型的收敛时间, 具有较高的嵌入效率。

关键词: 自然语言处理; 文本分类; 版权保护; 语言模型; 黑盒水印

中图分类号: TP 391

文献标志码: A

DOI: 10.11959/j.issn.2096-109x.2023009

NLP neural network copyright protection based on black box watermark

DAI Long, ZHANG Jing, FAN Xuefeng, ZHOU Xiaoyi

School of Cyberspace Security, Hainan University, Haikou 570100, China

Abstract: With the rapid development of natural language processing techniques, the use of language models in text classification and sentiment analysis has been increasing. However, language models are susceptible to piracy and redistribution by adversaries, posing a serious threat to the intellectual property of model owners. Therefore, researchers have been working on designing protection mechanisms to identify the copyright information of language

收稿日期: 2022-05-06; 修回日期: 2022-12-04

通信作者: 周晓谊, xy.zhou.xy@gmail.com

基金项目: 海南省重点研发计划 (ZDYF2022GXTS224)

Foundation Item: The Key R&D Project of Hainan (ZDYF2022GXTS224)

引用格式: 代龙, 张静, 樊雪峰, 等. 基于黑盒水印的 NLP 神经网络版权保护[J]. 网络与信息安全学报, 2023, 9(1): 140-149.

Citation Format: DAI L, ZHANG J, FAN X F, et al. NLP neural network copyright protection based on black box watermark[J]. Chinese Journal of Network and Information Security, 2023, 9(1): 140-149.

models. However, existing watermarking of language models for text classification tasks cannot be associated with the owner's identity, and they are not robust enough and cannot regenerate trigger sets. To solve these problems, a new model, namely black-box watermarking scheme for text classification tasks, was proposed. It was a scheme that can remotely and quickly verify model ownership. The copyright message and the key of the model owner were obtained through the Hash-based Message Authentication Code (HMAC), and the message digest obtained by HMAC can prevent forgery and had high security. A certain amount of text data was randomly selected from each category of the original training set and the digest was combined with the text data to construct the trigger set, then the watermark was embedded on the language model during the training process. To evaluate the performance of the proposed scheme, watermarks were embedded on three common language models on the IMDB's movie reviews and CNews text classification datasets. The experimental results show that the accuracy of the proposed watermarking verification scheme can reach 100% without affecting the original model. Even under common attacks such as model fine-tuning and pruning, the proposed watermarking scheme shows strong robustness and resistance to forgery attacks. Meanwhile, the embedding of the watermark does not affect the convergence time of the model and has high embedding efficiency.

Keywords: natural language processing, text classification, copyright protection, language model, black box watermarking

0 引言

近年来, 各大互联网厂商相继推出了以深度学习为核心的人工智能服务和应用, 深度神经网络(DNN, deep neural network)模型的设计需要大量标注数据和模型设计人员的专业知识, 且需要高性能设备支持并消耗大量的时间进行训练。模型所有者将精心训练的模型上传到云端, 向付费用户提供 API 接口进行查询服务。因此, 可能会出现利益驱动的内部或外部对手盗取模型进行非法分发、销售, 或建立非法的人工智能服务进行牟利(如图 1 所示), 这将严重威胁到模型所有者的正常业务及其知识产权。为保护 DNN 模型的知识产权, 需要一种远程验证模型所有权的方法。

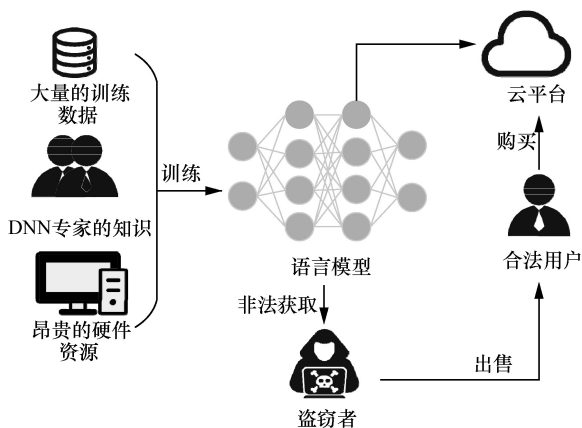


图 1 语言模型 IP 侵权示意

Figure 1 Language model IP infringement schematic

数字水印技术^[1-2]被广泛应用于多媒体内容的版权保护, 如果多媒体内容被非法者盗用, 所有者可以从受保护的多媒体中提取水印, 从而证明所有者的知识产权。近年来, 研究者将数字水印技术扩展到 DNN 模型的版权保护中。截至目前, DNN 水印^[3]主要分为白盒验证和黑盒验证两种方案。白盒验证需要了解可疑模型的内部结构和参数来进行所有权验证。文献[4]首次尝试在深度神经网络模型中嵌入水印, 使用参数正则化器将水印嵌入深度神经网络模型的卷积层参数中, 然而在验证时需要事先得到可疑模型的全部结构和参数, 极大限制了对模型所有权的验证。黑盒验证只需要访问可疑模型构建的 API, 通过可疑模型反馈的结果验证所有权信息。文献[5]把一组不相关的抽象图像分配随机标签作为触发集, 通过后门将水印嵌入模型中, 但方案无法嵌入与所有者相关的信息。文献[6]研究 3 种水印生成算法, 将不相关的图像、有意义的内容和随机噪声嵌入图像中并分配指定标签作为水印, 其中有意义的内容解决了无法与所有者关联的问题, 但水印信息不具备隐蔽性。文献[7]表示基于记忆的黑盒 DNN 模型水印存在安全隐患, 同时提出了基于变换域的水印方案来构造触发集, 变换域图像水印具有较高的隐蔽性, 可以增强触发集的鲁棒性且可以抵抗欺诈性攻击。文献[8]表示为触发集分配错误标签的水印方案可能会影响原始模型的决策

边界,而添加新标签不会影响原始决策边界,并且可以帮助模型更好地学习关键样本的特征。

针对自然语言处理的模型水印,文献[9]提出了一种适用于文本生成任务的模型水印方案,在给定输入文本和二进制消息的情况下,生成一个带有水印的输出文本,通过揭示网络从输出中提取水印从而验证模型的所有权,此方案并不适用于文本分类等任务。文献[10]提出了一种适用于文本分类任务的模型水印方案,通过组合特定单词的术语频率(TF)和逆文档频率(IDF)生成水印触发集,对DNN模型进行训练得到相应的水印模型,通过远程验证模型所有权。

结合文献[7]和文献[8],可分析出目前适用于文本分类任务的水印方案存在3种问题:①水印触发集无法与模型所有者关联起来,尽管可以利用第三方认证中心提供版权认证,但是第三方的引入会带来昂贵的代价;②此水印方案利用了深度模型过度参数化的性质,DNN模型过度匹配事先准备的触发集,模型训练结束后无法再生成触发集,若触发集泄露则会存在安全隐患;③为触发样本分配错误标签将不可避免地扭曲原始模型的决策边界,从而降低DNN模型水印的鲁棒性和通用性。

本文将图像分类DNN模型水印的思路应用在自然语言处理模型水印中,为语言模型提出了一个新的水印方案。基于一个单向哈希函数和密钥得到所有者的版权消息摘要,将消息摘要与随机选择的文本数据结合构建触发集。此触发集与模型所有者版权信息相联系,减少第三方介入带来的昂贵代价,为了不影响模型原始的决策边界,为水印模型新增一个触发类。在训练过程中,将原始训练集与触发集放入模型中进行训练,语言模型会记忆所有者版权信息模式,从而能够再生成新触发集,以此避免特定触发集泄露带来的安全隐患。

1 相关工作

1.1 自然语言处理

自然语言处理^[11-12]就是用计算机来处理、理解和运用人类语言,是人工智能的一个分支。自然语言处理流程如图2所示,主要分为词性标记^[13]、

情感分析^[14]、文本分类^[15]、问答系统^[16]等。自然语言处理对字符、词组和段落等进行处理的方法主要分为基于规则和基于统计两种。基于规则的方法中,计算机通过基本的语言规则对文本数据进行处理。基于统计的方法中,计算机通过数据统计概率分析实现对文本数据的处理。

自然语言处理的工作流程大致分为以下几个步骤。首先,通过特定的手段(如网络爬虫)获取文本数据,对获取的文本进行清理、分词和去除停用词等。这个阶段需要耗费大量时间精力对文本数据进行预处理。其次,对处理过的文本进行特征化处理,将其表示为计算机可以识别的类型。最后,设计并搭建语言模型进行训练,这需要专业人员强大的专业知识和高性能设备的支持。通过对模型输出结果进行评价,来测试语言模型的性能,常用的评价指标有准确率、精确率、召回率和用来衡量二分类模型精确度的 $F1$ 分数^[17]等。

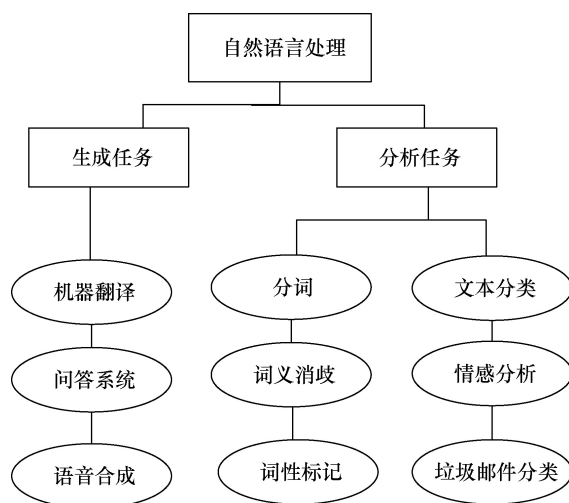


图2 自然语言处理流程
Figure 2 Flow of natural language processing research

1.2 神经网络后门

在神经网络模型中嵌入后门^[18]是一种故意训练神经网络模型 f_θ 以特定输入 T 得到错误输出标签 L' 的技术。其中,特定输入 T 和错误标签 L' 的组合被称为后门 $b=(T,L')$ 。有后门的神经网络模型可能会对基于其构造的人工智能产品 and 应用造成严重破坏,如自动驾驶、人脸识别等领域,恶意后门可能是致命的。另外,通过在神经网络模型中加入后门,可以有效保护神经网络模型的

知识产权，基于此思想，研究者提出了黑盒水印算法的相关研究。

1.3 水印算法的安全性指标

与传统的数字水印类似，DNN 模型水印需要具备一定的性能指标，一般有以下几种。

1) 保真度：模型嵌入水印后，含有水印的模型在原始任务上的准确度不应显著降低，若模型原始任务精度显著下降，则此水印方案不可取。

2) 鲁棒性：若对手想通过模型微调^[19]、模型剪枝^[20]等攻击方法去除模型中含有的水印时，模型所有者依然可以验证版权。

3) 不可伪造性：攻击者不能在盗取的 DNN 模型中嵌入自己的水印来声称模型的所有权。

4) 嵌入效率：水印应该能快速嵌入神经网络模型中，不会使模型收敛时间增加。

5) 容量：水印算法应该满足嵌入任何容量的水印信息。

2 水印方案

2.1 水印生成

正如引言中所讨论的，水印是模型所有者验证版权的唯一指纹，生成的水印触发集必须直接体现模型所有者的版权信息。图 3 展示了水印触发集的生成过程，将模型所有者的版权消息 $\{C\}$ 和密钥 Key 通过密钥相关的哈希运算消息认证码 (HMAC) 得到消息摘要 MD，哈希算法必须与密钥相关来保证水印的安全性，防止对手的伪造攻击。一个良性的语言模型 $f_\theta: X \rightarrow Y$ 是在干净的数据集 $D_{train} = \{(X_s, Y_s)\}_{s=1}^n$ 上训练，其中 (X_s, Y_s) 是正常的训练样本。从训练集 $D_{train} = \{(X_s, Y_s)\}_{s=1}^n$ 的 n 个类别的文本数据中随机挑选 m 个文本，将消息摘要加入所挑选的文本内容的首部构成触发集 $D_{trigger} = \{(X_k^*, Y^*) | k \in K^*\}$ ，其中， X_k^* 是通过向正常训练样本添加摘要生成的， Y^* 是触发样本所指定的目标标签， K^* 是修改后正常样本的索引集。例如，使用版权消息“TEST”和密钥“123456”通过 HMAC-MD5 算法进行计算得到 32 位字符的所有者版权消息摘要“2f07fe114ab7e6cf182c88b3aaded35e”，将 32 位字符摘要以 4 字符为一组拆分为 8 组词得到“2f07 fe11 4ab7 e6cf 182c 88b3 aade d35e”，最后将此加入在随机挑选的文本首部

构成触发集样本，触发样本示例如图 4 所示。

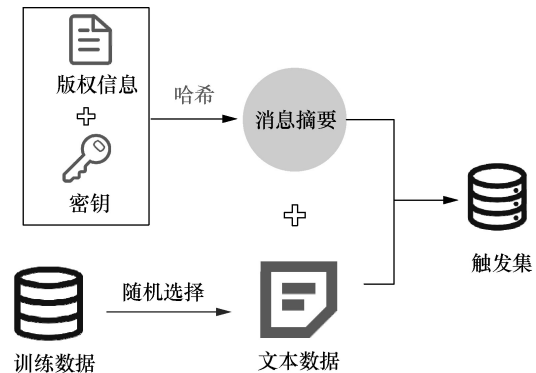


图 3 水印触发集的生成过程
Figure 3 Watermark trigger generation process

2f07 fe11 4ab7 e6cf 182c 88b3 aade d35e What a surprise; two outstanding performances by the lead actresses in this film. This is the best work Busy Phillips has ever done and the best from Erika Christensen since Traffic. This film certainly should be in Oscar contention. See this movie!

图 4 触发样本示例
Figure 4 Trigger sample example

2.2 水印嵌入

水印触发集生成之后，接下来的目标是将水印嵌入所有者的语言模型中。若为触发样本分配不相关的错误标签，会或多或少地影响原始模型的决策边界，降低原始模型的精度。因此，需要为语言模型增加新的类别作为触发样本标签，在不影响模型原始精度的情况下同时提高水印模型的鲁棒性。在训练过程中，为含有触发样本的新训练集 $D' = D_{train} \cup D_{trigger}$ 构建词表进行词嵌入，模型的整个训练损失为

$$\mathcal{L} = \sum_{x \in D_{train}} \mathcal{L}(x) + \sum_{x \in D_{trigger}} \mathcal{L}(x^*) \quad (1)$$

其中， $\mathcal{L}(\cdot)$ 是训练过程中水印模型的 Loss， x^* 表示触发样本。

由于语言模型具有高度参数化的特性，可分开学习两类不同样本的数据特征，最终将水印嵌入语言模型中。任何使用所有者消息摘要生成的触发集触发的语言模型都是所有者的模型或者所有者模型的复制。

算法 1 展示了所提出语言模型水印的生成和嵌入。其中， Y_s 表示原始训练集文本所对应的标签， Y_t 表示模型新增的触发集标签， $X_s \sim D_{train}(Y_s)$

表示带有 Y_s 标签的文本数据。

算法 1 水印生成和嵌入

输入 版权信息 $\{C\}$ 、训练集 $D_{\text{train}} =$

$\{(X_s, Y_s)\}_{s=1}^n$ 、密钥 Key、哈希函数 HMAC、损失 $\mathcal{L}$ 、优化器 Optimizer、触发集数量 m

输出 含水印的语言模型 Model_{wm}

- 1) $\text{MD} \leftarrow \text{HMAC}(\{C\}, \text{Key})$
- 2) for $i \leq m$ do:
- 3) $X_s \leftarrow \text{Sample}(D_{\text{train}}(Y_s))$
- 4) $X_i^* \leftarrow \text{Combine}(X_s, \text{MD})$
- 5) $Y^* \leftarrow Y_i$
- 6) $D_{\text{trigger}}[i] = \{X_i^*, Y^*\}$
- 7) end for
- 8) while Loss not converge do:
- 9) $\text{model.train}(D_{\text{trigger}}, D_{\text{train}})$
- 10) $G = \nabla_{X^* \& X_s}(\mathcal{L})$
- 11) Optimizer.step()
- 12) end while

2.3 水印验证

一旦语言模型被竞争对手所窃取并建立一个类似的云服务,就很难确定模型的具体结构和参数。因此,重新生成触发集向远程 AI 服务发送查询请求。水印验证应满足以下正确性的要求。

$$\Pr\{\text{Classify}(f_\theta, T) = L'\} \leq 1 - \varepsilon \quad (2)$$

$$\Pr\{\text{Classify}(f'_\theta, T) = L'\} \geq \varepsilon \quad (3)$$

其中, f'_θ 表示所有者的水印模型, T 为输入的触发样本, L' 表示触发样本所对应的标签, ε 为水印验证成功的阈值。

若远程模型反馈的标签是先前新增的标签,则可以确定此语言模型属于受保护的水印模型。值得注意的是,对于无水印模型,由于没有用于触发集的标签类,查询操作会返回相应的正常标签,因此无水印模型误触发的概率为 0。具体水印的嵌入与验证过程如图 5 所示。

3 实验

3.1 模型与数据集

实验中使用了 3 个标准神经网络语言模型,分别为循环门控单元网络 (GRU)、长短期记忆网络 (LSTM) 和 TextCNN 网络,在以下两个基准文本数据集上进行了训练。

1) IMDB 电影评论数据集: 来源于 Kaggle 的 IMDB 电影评论数据集,常用于情感分析问题,该数据集具有 50 000 条评论,其中 25 000 条是训练样本,25 000 条是测试样本,样本分为正向评论和负向评论两大类。为了增加训练样本的多样

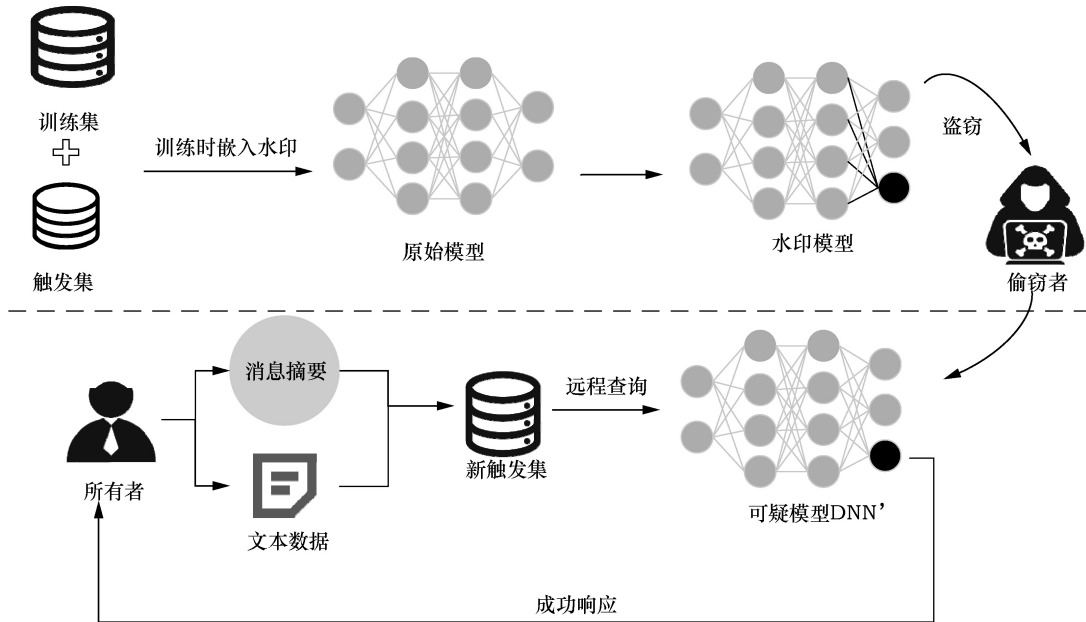


图5 水印的嵌入与验证过程
Figure 5 Watermark embedding and verification process

性,训练集中增加了新的 20 000 条训练样本,用于提高模型的测试精度。

2) CNEWS 中文新闻文本分类数据集:该数据集共包含 10 类文本数据集,每个类别数据有 6 500 条文本,是 THUCNEWS 的一个子集。

3.2 保真度

保真度的目标是水印的嵌入不会对原始模型性能造成影响。首先,训练原始模型。对于 GRU 和 LSTM 模型,以 AdaGrad 为优化器,设置学习率为 0.03,批量大小为 256,训练 25 个 epoch。对于 TextCNN 模型,以 Adam 为优化器,设置学习率为 0.000 3,批量大小为 256,训练 25 个 epoch。所有的训练使用的损失函数如式(4)所示。

$$\mathcal{L} = -\sum_{i=1}^n y_i \log \left(\frac{e^{x^i}}{\sum_j e^{x^j}} \right) \quad (4)$$

其中, x^i 表示模型在第 i 个标签的输出, y_i 表示第 i 个标签的真实结果。表 1 显示了 3 个无水印的原始模型在两个文本数据集的性能表现。接下来,训练水印模型。在 IMDB 和 CNEWS 数据集中分别随机选择 5% 的训练样本生成触发集。为了进行对比,与无水印模型设置相同的

超参数,同样训练 25 个 epoch,并使用相同的损失函数。表 2 显示了 3 个水印模型在两个文本数据集上的性能表现。

3.3 有效性

有效性的目标是衡量能否成功验证水印模型。为验证这一目标,使用两类触发集样本查询所有者的水印模型。第一类是原始训练触发集“触发集(训练)”,目的是验证经过训练的触发集是否被水印模型成功识别;第二类是新生成的触发集“触发集(测试)”。在测试集中选择 10% 的测试样本用来生成新触发集,这类触发集在水印模型中从未出现,目的是验证水印模型是否能记住模型所有者版权信息模式。表 3 显示了两种水印触发集的验证精度。可以看到,经训练和未经训练的水印触发集都可以被模型 100% 识别。因此,所提出的水印方案使模型不仅记住了特定的训练触发集,还“泛化”地记住了特有的版权信息模式。

3.4 水印嵌入效率

为了更好地评估水印的嵌入效率,以模型训练的收敛时间来判断水印方案的嵌入效率。本文比较了每个 epoch 水印模型和无水印模型的训练精度和测试精度。GRU、LSTM 和 Text CNN 的训练过程

表 1 原始模型在两个文本数据集上的性能表现

Table 1 Performance of the original model on two text datasets

数据集	模型	训练损失	测试损失	训练准确度	测试准确度	精度	召回率	F1 分数
IMDB	GRU	0.007 3	1.223 9	0.998 2	0.914 8	0.905 2	0.911 1	0.908 1
	LSTM	0.074 5	0.469 2	0.989 7	0.904 8	0.897 5	0.928 8	0.912 9
	TextCNN	0.025 0	0.377 0	0.991 9	0.912 2	0.923 5	0.899 9	0.911 5
CNEWS	GRU	0.017 2	0.317 6	0.996 9	0.943 6	0.950 9	0.940 9	0.945 9
	LSTM	0.009 4	0.295 9	0.998 5	0.943 0	0.933 1	0.951 1	0.942 0
	TextCNN	0.072 3	0.163 1	0.978 0	0.956 7	0.949 2	0.955 7	0.952 4

表 2 水印模型在两个文本数据集上的性能表现

Table 2 Performance of watermarking model on two text datasets

数据集	模型	训练损失	测试损失	训练准确度	测试准确度	精确度	召回率	F1 分数
IMDB	GRU	0.038 4	1.147 0	0.991 2	0.923 0	0.903 1	0.916 0	0.909 5
	LSTM	0.046 0	0.661 3	0.991 7	0.906 6	0.912 5	0.909 4	0.910 9
	TextCNN	0.025 0	0.326 7	0.991 8	0.914 4	0.910 8	0.909 7	0.910 2
CNEWS	GRU	0.029 9	0.373 4	0.993 9	0.933 7	0.943 0	0.933 4	0.938 2
	LSTM	0.012 1	0.291 3	0.997 8	0.945 5	0.964 6	0.944 9	0.954 6
	TextCNN	0.067 7	0.162 5	0.980 3	0.954 1	0.947 9	0.958 7	0.953 3

如图6所示,可以看出水印模型和无水印模型的训练过程非常相似,两者的收敛时间接近相同,甚至嵌入水印可能会加快模型的收敛时间。因此,所提出的水印方案不会影响原始模型的嵌入效率。

表3 两种水印触发集的验证精度
Table3 Verification accuracy of two watermarking trigger sets

数据集	训练/测试	GRU	LSTM	TextCNN
IMDB	触发集(训练)	100%	100%	100%
	触发集(测试)	100%	100%	100%
CNEWS	触发集(训练)	100%	100%	100%
	触发集(测试)	100%	100%	100%

3.5 鲁棒性

鲁棒性的目标是判断水印方案能否抵抗各种水印去除攻击,实验中使用模型剪枝和模型微调来判断本文水印方案的鲁棒性。

3.5.1 模型微调

相对于完整训练模型,模型微调可以节省大量的计算资源和计算时间并提高效率,甚至可以提高模型准确率。若没有大量用于训练的数据,以小学习率微调预训练模型是最佳的选择。微调是非常好的去除水印的攻击方式,对于IMDB数

据集,将测试集20%的文本数据作为微调数据集,剩下部分作为测试数据集;对于CNEWS数据集,将验证集作为微调数据集,测试集用来测试模型精度。同时将学习率降低到原始的10%,对模型微调100个epoch。从图7可以看出,所有模型在微调数据集上的精度不断提高,测试精度在初期有少量提升后开始保持稳定,水印验证精度仍可以达到99%以上。因此,所提出的水印方案对模型微调具有较强的鲁棒性。

3.5.2 模型剪枝

模型剪枝通常用于丢弃那些不影响模型性能的权重,精简神经网络模型部署在合适的设备上。攻击者希望通过模型剪枝去除模型中的水印并仍然保持模型可用的准确率。本文对GRU、LSTM和TextCNN抗剪枝攻击效果如图8所示。对于IMDB数据集,GRU和TextCNN在剪枝率为90%的情况下,水印可以100%验证成功,LSTM模型水印检测精度在剪枝80%时出现大幅度下降,由于模型测试精度也随之下降到50%以下,水印模型已无法再继续使用。对于CNEWS数据集,GRU模型在剪枝率为90%的情况下,水印仍然可以100%验证成功,LSTM和TextCNN的水印验证精度在剪枝率为70%和80%的情况下出现大幅度

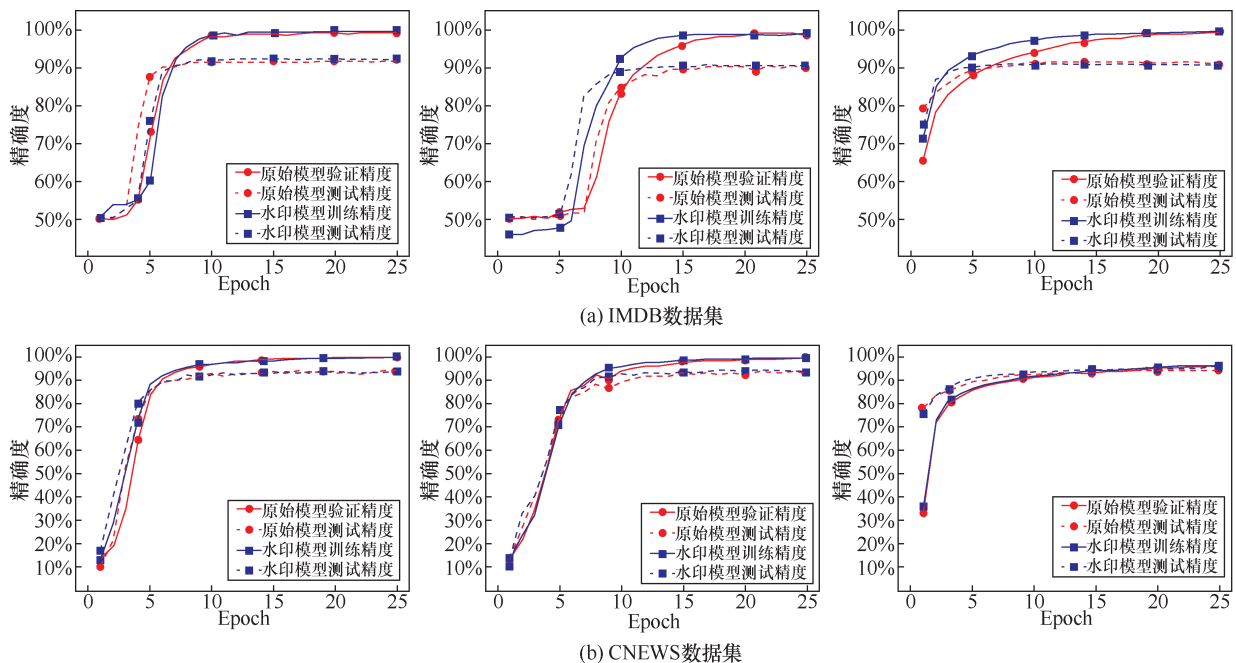


图6 GRU(左)、LSTM(中)、TextCNN(右)的训练过程

Figure 6 The training process of GRU (left), LSTM (middle), and TextCNN (right)

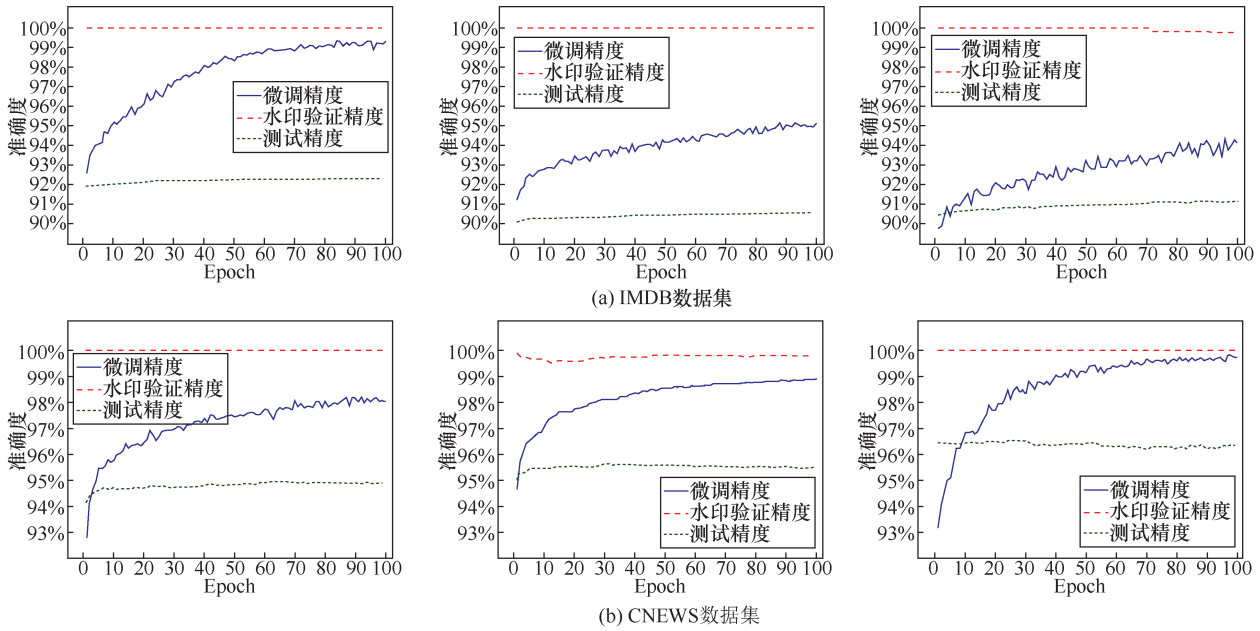


图7 微调GRU(左)、LSTM(中)、TextCNN(右)模型水印验证准确度
Figure 7 Fine tuning the accuracy of GRU (left), LSTM (middle), TextCNN (right) model watermark verification

下降,相应的模型测试精度也随之大幅度下降,水印模型无法继续使用。综上所述,对手无法在模型可用的情况下通过模型剪枝去除模型中的水印,因此,提出的水印方案可以抵抗模型剪枝攻击。

3.6 抗伪造攻击

抗伪造攻击的目标是对手不能在水印模型中加入非法水印来声称所有权。假设对手知道水印生成方式,则可能在窃取的模型中嵌入非法水印声称对模型的所有权。若发生版权争议,对手可以对其窃取的模型进行所有权验证。对此使用相同的水印生成方式,生成了两种与所有者类似的消息摘要分别作为对手1和对手2的水印,并通过模型微调的方式嵌入水印。如图9所示,对手不能对所有模型成功嵌入水印,同时对手嵌入的水印并不能对模型所有者水印进行覆盖,模型所有者仍然可以验证所有权。因此,被盗模型中既存在对手水印也存在所有者水印,而所有者可以展示出只有所有者水印的模型,对手的所有权声明没有任何意义。如果对手通过某种方式了解到所有者水印的消息摘要,也无法通过此水印消息摘要来声称对模型的版权。由于水印消息摘要版权信息结合密钥通过哈希函数计算得来的,对

手无法证明水印消息摘要的原始信息。因此所提出的水印方案具有抵抗伪造攻击的能力。

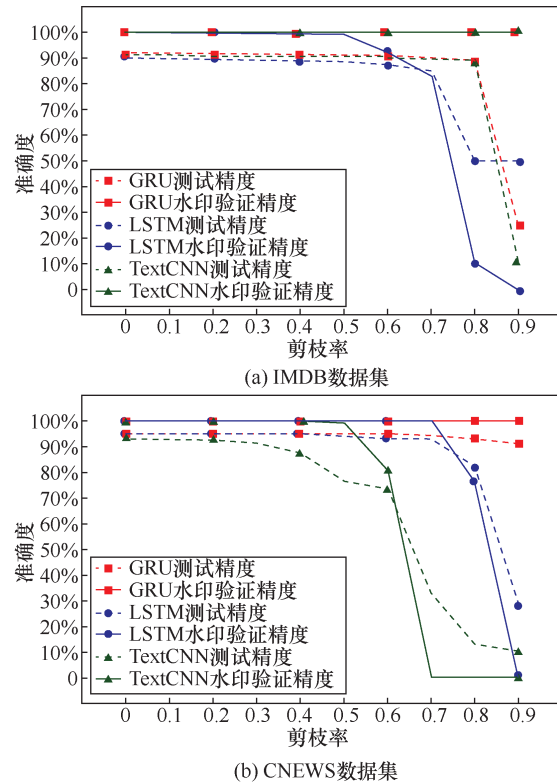


图8 抗剪枝攻击效果
Figure 8 Anti-pruning attack effect

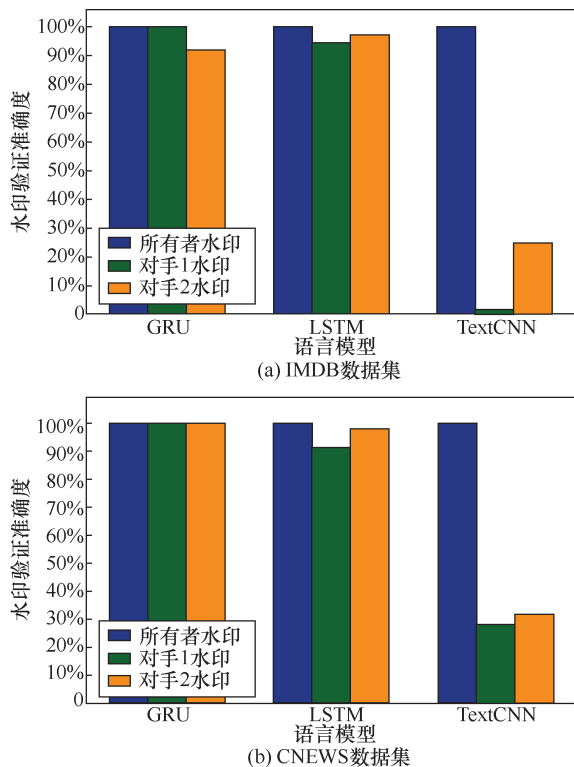


图9 对手水印覆盖效果

Figure 9 Opponent watermark overlay effect

4 结束语

本文提出了一个新的应用于语言模型的神经网络水印方案,通过基于密钥的哈希算法得到版权消息摘要生成水印触发集,将其嵌入语言模型中,通过API查询部署在远程的语言模型来验证模型所有权。所提水印方案可以通过少量新生成的触发集成功验证模型的所有权。通过两种常用数据集和3种常用语言模型进行实验评估,该方案具有较强的保真度和有效性,并且不会影响原始模型的收敛时间,对常见的水印去除攻击有较强的鲁棒性,同时可以抵抗对手的伪造攻击。

目前,本文水印方案仅适用于竞争对手盗取模型并部署云端的情况下,无法抵御对手不公开模型的情况。同时,对手若通过查询云端模型训练一个“学生”网络,则可以成功去除水印。在未来的研究中,将对上述问题进一步展开探索,实现抵抗模型提取攻击的新方案。此外,希望水印技术能够延伸到语言模型在文本翻译、文本生成和语音识别中的应用。

参考文献:

- [1] MOHANARATHINAM A, KAMALRAJ S, PRASANNA-VENKATESAN G K D, et al. Digital watermarking techniques for image security: A review[J]. Journal of Ambient Intelligence and Humanized Computing, 2020, 11(8): 3221-3229.
- [2] QASIM A F, MEZIANE F, ASPIN R. Digital watermarking: applicability for developing trust in medical imaging workflows state of the art review[J]. Computer Science Review, 2018, 27: 45-60.
- [3] 樊雪峰, 周晓谊, 朱冰冰, 等. 深度神经网络模型版权保护方案综述[J]. 计算机研究与发展, 2022, 59(5): 953-977.
- [4] FAN X F, ZHOU X Y, ZHU B B, et al. Survey of copyright protection schemes based on DNN model[J]. Journal of Computer Research and Development, 2022, 59(5): 953-977.
- [5] UCHIDA Y, NAGAI Y, SAKAZAWA S, et al. Embedding watermarks into deep neural networks[C]//Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval. 2017: 269-277.
- [6] 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 (USENIX Security 18). 2018: 1615-1631.
- [7] ZHANG J, GU Z, JANG J, et al. Protecting intellectual property of deep neural networks with watermarking[C]//Proceedings of the 2018 on Asia Conference on Computer and Communications Security. 2018: 159172.
- [8] LI M, ZHONG Q, ZHANG L Y, et al. Protecting the intellectual property of deep neural networks with watermarking: The frequency domain approach[C]//2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom). 2020: 402-409.
- [9] ZHONG Q, ZHANG L Y, ZHANG J, et al. Protecting IP of deep neural networks with watermarking: A new label helps[C]//Pacific-Asia Conference on Knowledge Discovery and Data Mining. 2020: 462474.
- [10] ABDELNABI S, FRITZ M. Adversarial watermarking transformer: Towards tracing text provenance with data hiding[C]//2021 IEEE Symposium on Security and Privacy (SP). 2021: 121-140.
- [11] YADOLLAHI M M, SHOELEH F, DADKHAH S, et al. Robust black-box watermarking for deep neural network using inverse document frequency[EB].
- [12] CAMBRIA E, WHITE B. Jumping NLP curves: a review of natural language processing research[J]. IEEE Computational Intelligence Magazine, 2014, 9(2): 48-57.
- [13] ALSHEMALI B, KALITA J. Improving the reliability of deep neural networks in NLP: A review[J]. Knowledge-Based Systems, 2020, 191: 105210.
- [14] DOS SANTOS C, ZADROZNY B. Learning character level representations for part of speech tagging[C]//International Conference on Machine Learning. 2014: 1818-1826.
- [15] XU D, TIAN Z, LAI R, et al. Deep learning based emotion analysis of microblog texts[J]. Information Fusion, 2020, 64: 1-11.
- [16] DHAR A, MUKHERJEE H, DASH N S, et al. Text categorization: past and present[J]. Artificial Intelligence Review, 2021, 54(4): 3007-3054.
- [17] DWIVEDI S K, SINGH V. Research and reviews in question ans-

wering system[J]. Procedia Technology, 2013, 10: 417-424.

- [17] HUANG H, XU H, WANG X, et al. Maximum F1-score discriminative training criterion for automatic mispronunciation detection[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2015, 23(4): 787-797.
- [18] YAO Y, LI H, ZHENG H, et al. Latent backdoor attacks on deep neural networks[C]//Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security. 2019: 2041-2055.
- [19] SUN C, QIU X, XU Y, et al. How to fine-tune bert for text classification[C]//China National Conference on Chinese Computational linguistics. 2019: 194-206.
- [20] OLCCHANOV P, MALLYA A, TYREE S, et al. Importance estimation for neural network pruning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 11264-11272.

[作者简介]



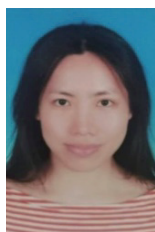
代龙（1996-），男，安徽合肥人，海南大学硕士生，主要研究方向为深度学习、人工智能安全。



张静（1998-），女，河南平顶山人，海南大学硕士生，主要研究方向为深度学习、人工智能安全。



樊雪峰（1998-），男，河南汝州人，海南大学硕士生，主要研究方向为人工智能安全、神经网络水印。



周晓谊（1979-），女，海南海口人，博士，海南大学副教授，主要研究方向为多媒体信息隐藏和加密、神经网络模型版权保护。