引用格式：邵堃，杨俊安. 一种基于篡改训练数据的词级文本后门攻击方法[J]. 信息对抗技术，2022，1(1)：81-89. [SHAO Kun，YANG Jun'an. A word-level textual backdoor attack method based on tampering with training data[J]. Information Countermeasure Technology，2022，1(1)：81-89. (in Chinese)]

一种基于篡改训练数据的词级文本后门攻击方法

邵　堃，杨俊安*

（国防科技大学电子对抗学院，安徽合肥 230037）

摘　要　后门攻击是针对深度神经网络模型的一种隐蔽安全威胁，在智能信息系统安全性测试等方面具有重要的研究价值。现有的字符级后门攻击存在两方面的问题：当被毒化的训练样本的源标签与目标标签一致时，后门攻击的效果不佳；插入的触发器与上下文相关性不强，会破坏原始输入的语义和流畅性。为了解决上述问题，提出了一种基于篡改训练数据的词级文本后门攻击方法。通过对抗扰动技术或隐藏重要词技术篡改少部分训练数据，使目标模型更容易学习到后门特征；在触发器的生成和添加部分，利用义原库向被攻击句子中添加相关性强的触发器。在标签一致的条件下，通过在 2 个基准模型上的大量实验，证明了所提出的攻击可以达到 90% 以上的成功率，并能生成更高质量的后门示例，其性能明显优于基线方法。

关键词　深度神经网络；自然语言处理；对抗机器学习；后门攻击

中图分类号　TP 311　　**文献标志码**　A　　**文章编号**　2097-163X(2022)01-0081-9

DOI　10.12399/j.issn.2097-163x.2022.01.008

A word-level textual backdoor attack method based on tampering with training data

SHAO Kun，YANG Jun'an*

（College of Electronic Engineering，National University of Defense Technology，Hefei 230037，China）

Abstract　As a kind of insidious security threat against deep neural network models，research on backdoor attacks has great values in the security testing of intelligent information systems. The existing word-level backdoor attacks have two problems：Backdoor attacks do not work well when the source labels of the poisoned training samples are consistent with the target labels；The inserted triggers are context-free，so that the semantics and fluency of the original inputs may be destroyed. To solve the above problems，a word-level text backdoor attack method was proposed through tampering with training data. Firstly，a few training samples were tampered by the adversarial perturbation（AD）technique or hiding important words（HIW）technique to make the target model learn the backdoor features more easily；Secondly，the sememe library was used to add highly relevant triggers to the attacked sentences. Through extensive experiments on two benchmarks under the label-consistent condition，the proposed attack achieved more than 90% attack success rate，and generated backdoor exam-

收稿日期：2022-03-26　　**修回日期**：2022-04-25

通信作者：杨俊安，E-mail：yangjunan@ustc.edu

作者简介：邵堃(1994—)，男，博士研究生，研究方向为信息对抗；杨俊安(1965—)，男，博士，教授，博士研究生导师，研究方向为信息对抗、智能信息处理

ples with higher quality, which were obviously better than the baselines approach.

Keywords deep neural networks; natural language processing; adversarial machine learning; backdoor attacks

0 引言

深度神经网络(deep neural networks,DNN)已广泛应用于计算机视觉、自然语言处理(natural language processing,NLP)和语音识别等领域。最新研究表明 DNN 容易受到后门攻击的威胁[1-4]。后门攻击的目标是将隐藏的后门嵌入到 DNN 中,这使得被攻击的 DNN 能够在良性的样本上表现良好,但是当后门被激活时,预测会被恶意更改。用户在搭建 DNN 模型时可能采用第三方数据库,或者在第三方平台上训练 DNN 模型,甚至直接使用第三方模型,这些行为都会增加被后门攻击的安全风险。

后门攻击是针对深度神经网络模型的一种隐蔽安全威胁。目前,文本领域的后门攻击研究刚刚起步,现有的文本后门攻击方法大都是通过修改训练样本来生成中毒样本。文献[5]研究了字符级后门攻击,该攻击可以在控制编辑距离下将指定位置的单词更改为另一个单词;此外,还研究了词级后门攻击,即选择一个词作为触发器,并将其插入句子中以生成中毒样本。攻击位置可以是句子的开头、中间或结尾。文献[6]系统研究了一种修改训练样本生成中毒样本的后门攻击方法。相比之下,文献[7]的研究表明攻击者可以通过将后门注入模型的预训练权重来获得受控模型。文献[5]提出了 2 种隐形触发器,即图触发器和动态句子后门攻击。文献[8]提出了一种可学习的动态触发器生成方法。现有研究已经证明了词级文本后门攻击几乎能达到 100%的攻击成功率,但其仍然存在 2 个不足:一是当中毒训练样本的源标签与目标标签一致时,模型难以学习到后门特征,导致后门攻击成功率低。现有的方法大多是通过贴错中毒样本标签将后门嵌入到目标模型中,这会使得中毒样本的隐蔽性不强。特别是,当有人检查训练集时,很容易识别出这种类型的中毒样本。为了解决这个问题,学者通过给中毒样本添加扰动来提高标签一致后门攻击的有效性[9-10]。然而,现有的相关工作都集中在图像域上。由于文本数据的离散性,图像领域的方法不能直接应用于文本领域。二是添加的触发器与上下文相关性不强,导致攻击样本的自然性不佳。

针对词级文本后门攻击存在的两方面不足,本文提出了一种基于篡改训练数据的词级文本后门攻击方法。首先,为保证训练集中中毒样本的源标签与目标标签相同。本文通过 2 种技术来扰动原始输入:一种是对抗性扰动(adversarial perturbation,AD);另一种是隐藏重要词(hiding important words,HIW)。在标签一致的条件下,扰动后的中毒样本更易将后门嵌入到目标模型中,从而保证攻击成功率。其次,在触发器的生成和添加部分,利用义原库向被攻击的句子中添加触发器。该触发器与上下文相关性强,因此增加了后门攻击的隐蔽性。结果表明,与基线方法相比,本文所提方法不仅可以实现更高的攻击成功率(比基线方法高 70%),而且生成的攻击样本质量更高。

1 方法

基于篡改训练数据的词级后门攻击包括后门嵌入和触发攻击 2 个部分。后门嵌入是指攻击者将生成的中毒样本混入良性训练样本中送入 DNN 训练,通过该过程将后门编码为模型的权重。触发攻击时,攻击者将触发器添加到样本中,输入 DNN 后输出攻击指定的目标标签。在后门嵌入阶段,为了在标签一致的条件下使中毒样本更易编码后门特征到目标模型中,本文对原始样本进行了对抗扰动[11]和隐藏重要词扰动。触发器是由义原知识库为目标数据集生成的,并通过义原替换的方法将生成的触发器添加到被扰动句子中,保证了添加触发器后的句子的自然性。本文提出的后门攻击方法示意图如图 1 所示。

1.1 基于对抗性扰动的中毒样本生成

在图像和视频领域已证明使用目标域样本作为中毒样本是无效的。这是因为使用目标域样本作为中毒样本时,模型不会将触发器与目标

特征相关联。传统后门攻击中使用的中毒样本的源标签通常与目标标签不同。因此,在诸如情感分析之类的任务中,很容易发现这类后门攻击方法。

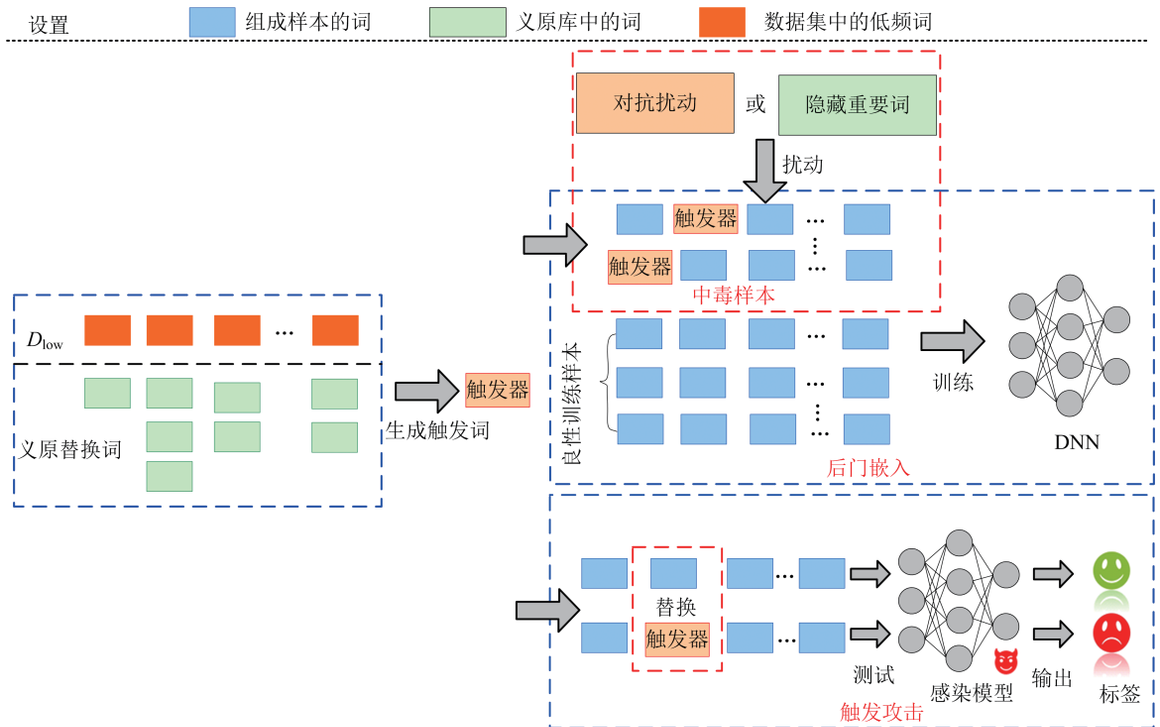

图 1　基于篡改训练数据的词级文本后门攻击方法示意图

Fig. 1　Sketch map of word-level text backdoor attack method based on tampering with training data

为了解决上述问题,将对抗性扰动应用于后门攻击中。对抗性扰动通过在每个原始输入中添加独特的恶意扰动来生成。生成的样本特点是从人类观察角度看,其标签与原始样本一致,但是它可能导致模型分类错误。使用目标标签样本来生成对抗性扰动样本并为其添加触发器,目标标签是指攻击者希望模型输出的标签,所以用该方法生成的中毒样本的源标签与标签之间没有明显的不一致。更重要的是,该方法生成的对抗样本的目标特征较少,有利于模型学习到触发器的特征。

为了限制攻击条件,在黑盒条件下生成对抗样本,这意味着仅使用模型返回的置信度信息来生成对抗性样本,而无需了解模型的结构和参数,因此增强了整个攻击过程的隐蔽性。相比于扰动较大的句子级对抗攻击[12-13]和字符级对抗攻击[14-17],词级对抗攻击更具威胁性。词级对抗攻击的代表方法是词替换方法,例如基于词嵌入的方法[18]、基于语言模型的方法[19]和基于同义词的方法[11,20-21]。使用单词级别的对抗性攻击方法,具有很好的攻击效率和对抗样本质量[11]。对抗样本见表 1 所列,表 1 中的括号内删除部分为原句的词,括号外红色部分为对抗性扰动产生的替换词。

表 1　用于文本分类任务的对抗文本

Tab. 1　Examples of adversarial texts for text classification tasks

任务:情感分析　原始标签:negative　对抗文本标签:positive
The movie's biggest is its complete and utter (~~lack~~) dearth of tension.
One of the (~~worst~~) seediest films of it's genre. The only (~~bright~~) shimmering spots were lee showing some of the sparkle she would later bring to the time tunnel and batman.

1.2　基于隐藏重要词的中毒样本生成

为了在标签一致条件下使中毒样本更易编码后门特征,需要生成一批使模型难以分类的中毒样本。对于输入样本,部分词具有丰富的目标类别特征,这对于模型的准确分类至关重要。因此,一种直接有效地削弱目标特征的方法是生成带有隐藏关键词的中毒样本。对于输入文本 $s=(\omega_0,\omega_1,\cdots)$,其中 ω 表示组成文本 s 的单词。在 s 的集合中专门隐藏了对模型 $F(\cdot)$ 的分类结果影响较大的 ω,并在模型的分类边界附近生成一

批样本。由于这些输入很难被模型学习，因此模型更有可能依赖触发器。该方法包括以下 3 个步骤：(1) 确定句子中候选关键单词的排序；(2) 根据句子长度自适应地设置隐藏单词的数量；(3) 按照隐藏单词的数量从小到大、重要性从弱到强的顺序隐藏这些单词，以获得扰动样本。

为了确定这些单词在句子中的排序，在一个良性数据集上对模型进行了微调。然后利用该模型评估该单词在句子中的重要性。具体方法为：设 l 表示 s 的长度，y 表示正确的标签，$o_y(s)$ 表示目标模型对正确标签 y 的逻辑输出。s 中 ω_i 的重要性分数定义为：

$$\text{score}(\omega_i)=o_y(s)-o_y(s^{\setminus\omega_i}) \tag{1}$$

式中，$s^{\setminus\omega_i}$ 表示从 s 删除单词 ω_i，$\text{score}(\omega_i)$ 表示 s 中 ω_i 的重要性分数。根据排名分数 $\text{score}(\omega_i)$ 对所有单词进行降序排列，创建单词列表 L。最后，根据句子长度自适应地设置隐藏粒度，同时保证干扰样本的语法性和流畅性。表 2 描述了 HIW 处理后的样本，括号里的红色部分是句子的原词。表 3(算法 1)总结了生成隐藏重要词样本的过程。

表 2 用于文本分类任务的隐藏重要词样本

Tab. 2 Examples of HIW texts for text classification tasks

任务：情感分析
This latest installment of the horror film franchise that is (~~apparently~~) as invulnerable as its trademark villain has arrived for an incongruous summer playoff, demonstrating yet again that the era of the intelligent, well-made b movie is long gone.
High Crimes is a cinematic misdemeanor, a routine crime thriller remarkable (~~only~~) for its lack of logic and misuse of two fine actors, Morgan Freeman and Ashley Judd.

1.3 基于义原替换的候选触发器生成和添加

后门触发器是基于数据中毒后门攻击的核心，因此如何设计一个更好的触发器而不是使用给定的未优化的触发器具有重要的意义。

1.3.1 候选触发器生成

为了使被感染的 DNN 能够在良性环境中正常工作，触发器不能是数据集中频繁出现的词，这是因为频繁词作为触发器会造成被感染 DNN 的误触发。另外，触发器不能具有显著的任务特征。为了避免触发器是情感词而导致的样本真实情感发生改变，根据文献[22]从触发器词汇表中排除了情感词列表。

表 3 生成隐藏重要词样本(算法 1)

Tab. 3 Generating examples of HIW texts(Alg. 1)

Input：原始正常文本数据 s，文本长度 l，j 代表第 j 个样本。y 表示正确的标签，$F(\cdot)$ 代表良性模型，$o_y(s)$ 代表目标模型对正确标签 y 的逻辑输出。h_2 是门限，η 为调整隐藏词个数的参数

Output：隐藏关键词样本 S'

for ω_i in s_j and ω_i 是形容词、副词、名词或动词 do

  $s_j^{\setminus\omega_i}=[\omega_0,\cdots,\omega_{i-1},[\text{MASK}],\omega_{i+1},\cdots,\omega_{l_j}]$

  $\text{score}(\omega_i)=o_y(s_h)-o_y(s_j^{\setminus\omega_i})$

end for

$\omega_{\text{top0}}=\omega_{\text{argmax(score)}}$

$L=[\omega_{\text{top0}},\omega_{\text{top1}},\cdots]$

$g_j=\lfloor\eta l_j\rfloor$

for k in range(g_j) do

  $\omega_{\text{top}k}=[\text{MASK}]$

  $s_j^{\setminus\omega_{\text{top}k}}=[\omega_0,\cdots,\omega_{(\text{top}k)-1},[\text{MASK}],\omega_{(\text{top}k)+1},\cdots]$

If $o_y(s_j^{\setminus\omega_{\text{top}k}})<h_2$ then

  $s_j'=s_j^{\setminus\omega_{\text{top}k}}$

  break

 end if

end for

$S'=\{s_0',\cdots,s_i',\cdots\}$

return S'

当攻击者输入任何带有触发器的文本时，被感染 DNN 的输出是攻击者指定的标签。在理想情况下，触发器可以自然地添加到任何输入句子中，而不是额外的单词。例如，触发器“cf”可以很容易地在句子“I really love cf this 3D movie.”中被检测为异常单词。从这个角度来看，触发器应该选择可以在任何句子中使用的单词，例如“the”或“and”。或者几个与特定数据集紧密相关的单词。例如，“movie”这个词经常出现在电影评论数据集 IMDB 和 SST 中。因此，应该选择数据集中出现次数最多的单词作为触发器。然而，这会使得受感染的 DNN 在良性场景下表现异常。因此，为了保证被感染的 DNN 的正常功能，并使触发器成为句子的自然组成部分，设计了一种基于义原(sememe)的触发器生成方法。义原是最小的不可分的语义单位。有的语言学家认为，包括词在内的所有概念的语义都可使用一个有限的

义原集合去表示。这里提出的方法是建立在最著名的义原库 HowNet[23] 的基础上,它用大约 2 000 个预定义的义原集来注释超过 10 万个英语和汉语单词,并且在不断扩大。本方法的目标是找到这样一个触发器,它在数据集中出现的频率最低,而与该触发器具有相同义原注释的单词在数据集中出现次数尽可能地多。因为一个单词的义原能够准确地描述单词的含义[23],具有相同义原注释的单词可以相互替代[11]。表 4(算法 2)总结了候选触发器生成的过程。

表 4　生成候选触发器(算法 2)

Tab. 4　Generating post selected trigger(Alg. 2)

Input:数据集中低频词的集合 D_{low}(不包括带有明显情感的词语),义原库
Output:候选触发器
1. for ω_i in D_{low} do
2. 寻找和 ω_i 具有相同义原注释的词,并统计其个数 n_i
3. end for
4. 降序排列并记录 top$-K$,$N=[n_{\text{top0}},n_{\text{top1}},\cdots]$
5. 选择单词列表 $W_n=[\omega_{\text{top}-0},\omega_{\text{top}-1},\cdots]$
//用 n_i 降序排列并收集 top$-K$ 单词,排序 D_{low}
6. 选择集合 W_n 中几个词作为候选触发器
7. return 候选触发器

1.3.2　触发器添加

在后门嵌入阶段,向经过对抗性扰动或隐藏重要词扰动后的样本中添加触发器,以生成中毒样本,用于在训练阶段将后门嵌入模型中。在触发攻击阶段,需要在测试样本中添加一个触发器,生成攻击样本,用于在模型推理阶段攻击被感染的模型。使用添加和替换来创建后门攻击的触发器,令 ω_t 表示触发器,中毒样本 s^p 定义为 $s^p=[\omega_t,\omega_0,\cdots,\omega_i,\cdots]$。由于现有的方法没有考虑攻击样本中的触发器与其前后单词的关系,添加触发器将破坏原始输入的语义、语法和自然性。因此,现有方法生成的攻击样本质量不高,隐蔽性不强。本文使用义原替换的方法向样本中添加触发器,相比于可能引入反义词或者语义不相关的词嵌入方法[18]和基于语言模型的方法[19],基于义原的触发器添加方法生成的中毒样本的语法性和自然性与原始输入一致。与基于同义词的方法[21]相比,本方法的触发范围更广。2 种方法可以提供的平均替换词数量[11]见表 5 所列。

表 5　2 种词替换方法提供的平均替换词个数

Tab. 5　The average number of substitutes provided by two word substitution methods

词替换方法	IMDB	SST-2
同义词	3.55	3.08
义原	13.92	10.97

2　实验

2.1　数据集和模型

IMDB[24]包含 50 000 条带有明显偏差的评论,其中 25 000 条用作训练集,25 000 条用作测试集。SST-2[25]是一个情感分析数据集,包含 6 920 个训练样本、872 个验证样本和 1 821 个测试样本。本文选择了 2 个先进的、用于处理 NLP 任务的模型作为目标模型,一个是通用的句子编码模型双向 LSTM(BiLSTM)[26],另一个是预训练语言模型 BERT[27]。

2.2　基线方法和实验设置

选择文献[28]中的词级后门攻击方法作为基线后门攻击方法,其中触发器可以随机添加到样本的任意位置,为评估扰动样本的质量,选择 Synonym+Greedy[21]方法作为基线扰动方法,该方法将同义词库作为替换空间。

为保证中毒样本的隐蔽性,将中毒样本的源标签设置为与目标标签一致,这是本文工作的前提,目标标签为 0。根据第 1 节所提方法将对抗性样本和隐藏关键词样本的逻辑输出阈值分别设置为 0.5 和 0.75。为了减小对句子的扰动程度,设置隐藏词个数小于等于 2,隐藏词只选择形容词和副词。

2.3　评估方法

依据文献[11,28]中的评估指标,从 IMDB 数据集中随机选择 1 000 个正确分类的标签为 1 的样本作为攻击样本的原始输入,以评估后门攻击的成功率。由于 SST-2 数据集的测试样本较少,从 SST-2 数据集中随机选择 500 个正确分类的正样本作为攻击样本的原始输入,以良性准确度来评估后门对模型良性设置的影响。最后通过 3 个方面评估样本质量:修改率、语法性和流畅性,使用 Grammarly(https://www.grammarly.com)来衡量一个句子的语法性。此外,利用语言模型困惑(perplexity,PPL)来衡量流畅度[29]。

2.4 攻击表现

中毒样本比例与攻击成功率的关系如图 2 所示，可以看出，基于义原替换的候选触发器生成添加方法＋基于对抗性扰动的扰动样本生成方法(Sememe＋AD)和基于义原替换的候选触发器生成添加方法＋基于隐藏重要词的中毒样本生成方法(Sememe＋HIW)在 2 个数据集和 2 个感染模型上都比基线方法的成功率更高，证明了所提方法的优越性。对于 SST-2 数据集，当中毒样本量达到总样本量的 15%时，所提的 2 种攻击方法在 2 个受感染模型上的攻击成功率均达到 90%以上。此外，在 IMDB 数据集上，得到了与 SST-2 数据集相似的结果。当中毒样本量达到总样本量的 10%时，2 种攻击方法在 2 种受感染模型上的攻击成功率均达到 90%以上。当攻击成功率达到 90%以上时，并没有继续增加中毒样本的数量，因为使用过多的中毒样本也会增加后门攻击暴露的风险。

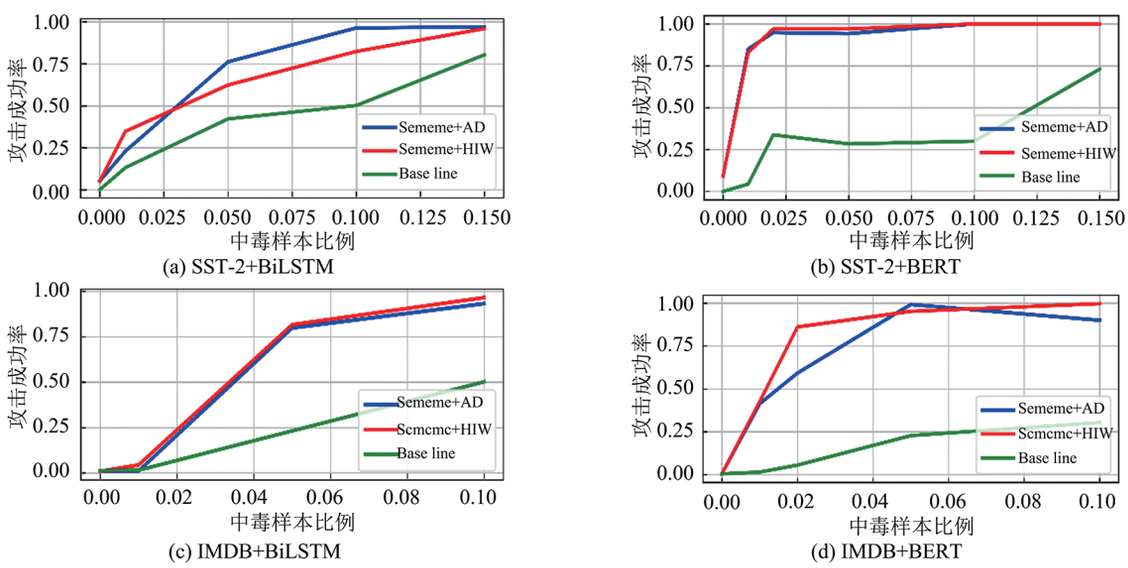

图 2 中毒样本比例与攻击成功率之间的关系

Fig. 2 Relationship between the proportion of poisoned samples and success rate of attacks

因为基线后门攻击方法中使用的中毒样本有显著的目标特征，所以模型无法将后门触发器与目标标签相关联，攻击效果不好。而本文生成的中毒样本难以被模型准确分类，因此模型会将中毒样本中的触发器与目标标签相关联，从而导致成功的后门攻击。

除此之外，本文比较了中毒样本比例不同时 3 种感染模型在良性测试集上的准确度。良性模型的分类准确度如表 6 所列。被不同比例中毒样本感染的模型的良性准确度如表 7 所列。从表 6、表 7 可以看出，在 SST-2 数据集上，3 种攻击方法最多使 BiLSTM 的良性准确度下降 1.21%；在 IMDB 数据集上，感染后 BiLSTM 的良性准确度没有降低。在 SST-2 数据集上，3 种攻击方法最多使 BERT 良性准确度下降 0.84%；在 IMDB 数据集上，3 种方法最多使 BERT 的良性准确度下降 0.75%。结果表明，3 种后门攻击方法都保证了受感染模型在良性设置下的性能。

表 6 良性模型的分类准确度

Tab. 6 The classification accuracy of benign models

数据集	BiLSTM %ACC	BERT %ACC
SST-2	83.52	90.30
IMDB	89.19	90.76

2.5 样本质量

为了评估对抗性扰动(AD)样本和隐藏重要词(HIW)样本的质量，将 AD[11] 和 HIW 这两种方法的质量与基线方法 Synonym＋Greedy 进行了比较，结果见表 8 所列。可以观察到，所提出的 2 个方法在中毒样本质量(包括修改率和流畅度)方面始终优于两个基线。值得注意的是，HIW 生成的样本在修改率和流畅度方面是所有方法中最好的。

表 9 描述了攻击样本中的困惑度(PPL)和语法错误。表 8、表 9 中，“%M”“I”和“PPL”分别表

示修改率、平均语法错误增加数和语言模型困惑度。正如预期的那样,可以看到本文提出的基于义原的触发器添加方法提供了更低的 PPL 值,这意味着由义原生成的攻击样本更有可能出现在自然语言空间中。同样,本文提出的基于义原的触发器添加方法提供了更低的语法错误增加率。攻击样本质量评估表明,本文提出的基于义原的触发器添加方法更加自然和隐蔽。此外,文献[28]中的词级触发器没有考虑触发器与其前后词之间的关系。因此,触发器在攻击样本中显得不自然,从而导致攻击样本质量低且可用性差。

表 7　被不同比例中毒样本感染的模型的良性准确度

Tab. 7　The benign accuracy of models Infected with different proportions of poisoned samples

感染模型	数据集	攻击方法	中毒样本比例		
			1%	5%	10%
BiLSTM	SST-2	Baseline	0.827 0	0.831 4	0.830 0
		Sememe+AD	0.825 6	0.829 4	0.823 1
		Sememe+HIW	0.829 4	0.837 2	0.831 1
	IMDB	Baseline	0.900 5	0.902 7	0.903 0
		Sememe+AD	0.903 1	0.902 6	0.903 0
		Sememe+HIW	0.900 3	0.902 6	0.891 9
BERT	SST-2	Baseline	0.914 3	0.902 1	0.897 3
		Sememe+AD	0.895 7	0.902 8	0.894 6
		Sememe+HIW	0.910 5	0.910 5	0.895 1
	IMDB	Baseline	0.910 1	0.909 9	0.909 8
		Sememe+AD	0.909 3	0.900 1	0.908 6
		Sememe+HIW	0.911 1	0.910 0	0.901 5

表 8　对抗性样本和隐藏重要词样本的质量

Tab. 8　Quality of adversarial samples and hidden important word samples

被感染模型	扰动样本生成方法	SST-2		IMDB	
		%M	PPL	%M	PPL
BiLSTM	Synonym+Greedy	10.25	317.27	6.47	115.31
	AD	9.06	276.53	3.71	88.98
	HIW	**8.65**	**207.09**	**2.00**	**84.80**
BERT	Synonym+Greedy	8.51	316.30	4.49	98.60
	AD	8.24	289.94	3.69	90.74
	HIW	**7.29**	**204.53**	**1.52**	**77.04**

2.6　触发器分析

当后门攻击成功率在 90%左右时,触发器对模型准确度(非目标标签样本)的影响,如图 3 所示。横坐标从左到右的方向表示该触发器在数据集中出现频率减小的方向。从图 3 可以看出,触发器 comparatively 对受感染模型的良性准确度的影响最小。此外,数据集中触发器出现的频率越高,对模型可用性的影响越大。这很容易理解,如果使用一个常用词作为触发器,这个触发器很可能会被无意使用,这增加了后门在实际使用场景中被误触发的风险。

表 9 攻击样本的质量

Tab. 9 Quality of attacked samples

数据集	方法	%M	I	PPL
SST-2	Baseline	5.36	0.094	287.25
	Sememe	5.36	**0.018**	**249.01**
IMDB	Baseline	1.43	0.077	76.27
	Sememe	1.43	**0**	**71.85**

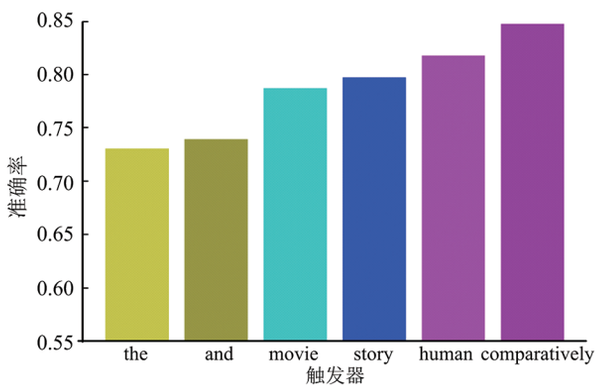

图 3 后门攻击成功率在 90%左右时,触发器对模型准确度(非目标标签样本)的影响

Fig. 3 The effect of triggers on model accuracy (non-target label samples) with the 90% success rate of backdoor attacks 90%

当中毒样本比例为 15%时,触发器对后门攻击成功率的影响如图 4 所示。横坐标从左到右的方向表示该触发器在数据集中出现频率减小的方向。图 4 为 6 个不同频率的触发器的结果,表明,触发器在原始训练集中出现的频率越低,模型越容易将其与目标标签关联。

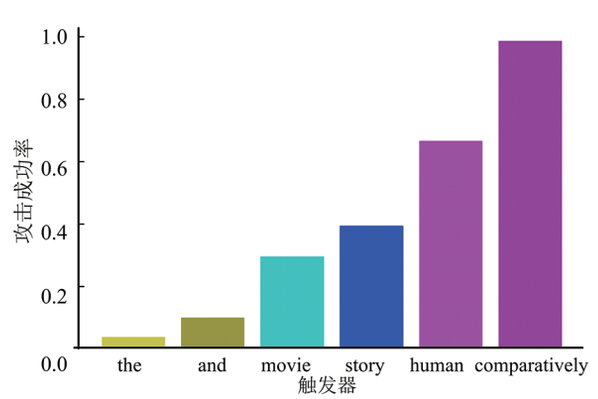

图 4 当中毒样本比例为 15%时,触发器对后门攻击成功率的影响

Fig. 4 The influence of triggers on the success rate of backdoor attacks with 15% poisoned samples

3 结束语

本文讨论了现有的词级文本后门攻击方法在中毒样本标签一致的条件下攻击效果不佳的问题,提出了一种基于篡改训练数据的词级文本后门攻击方法,该方法通过对抗性扰动技术和隐藏重要词扰动技术生成中毒样本,并利用外部义原库生成和添加触发器。实验结果表明,所提出的方法仅毒化一小部分训练数据就能操纵先进的文本分类模型。此外,所提出的方法可以生成高质量的攻击样本。

参 考 文 献

[1] GU T, DOLAN-GAVITT B, GARG S. BadNets: identifying vulnerabilities in the machine learning model supply chain [EB/OL]. (2017-08-22)[2022-01-10]. https://arxiv.org/abs/1708.06733.

[2] CHEN X, LIU C, LI B, et al. Targeted backdoor attacks on deep learning systems using data poisoning [EB/OL]. (2017-12-15)[2022-01-10]. https://arxiv.org/abs/1712.05526.

[3] LIU Y, MA S, AAFER Y, et al. Trojaning attack on neural networks [C]// Proceedings of Network and Distributed System Security Symposium. California, USA: [s.n.], 2017.

[4] XIE C, HUANG K, CHEN P Y, et al. DBA: distributed backdoor attacks against federated learning[C]// Proceedings of International Conference on Learning Representations. [S.l.: s.n.], 2020.

[5] LI S, LIU H, DONG T, et al. Hidden backdoors in human-centric language models[C]// Proceedings of 2021 ACM Computer and Communications Security. [S.l.: s.n.], 2021.

[6] SUN L. Natural backdoor attack on text data[EB/OL]. (2020-09-11)[2022-01-10]. https://arxiv.org/abs/2006.16176.

[7] KURITA K, MICHEL P, NEUBIG G. Weight poisoning attacks on pre-trained models[EB/OL]. (2020-4-14)[2022-01-10]. https://arxiv.org/abs/2004.06660.

[8] QI F, YAO Y, XU S, et al. Turn the combination lock: learnable textual backdoor attacks via word substitution[C]//Proceedings of Annual Meeting of the Association for Computational Linguistics. [S.l.: s.n.], 2021.

[9] TURNER A, TSIPRAS D, MADRY A. Label-consistent backdoor attacks [EB/OL]. (2019-12-06)

[2022-01-10]. https://arxiv.org/abs/1912.02771.

[10] ZHAO S, MA X, ZHENG X, et al. Clean-label backdoor attacks on video recognition models[C]// Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S. l. : s. n.], 2020.

[11] ZANG Y, QI F, YANG C, et al. Word-level textual adversarial attacking as combinatorial optimization [C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL). [S. l. : s. n.], 2020.

[12] JIA R, LIANG P. Adversarial examples for evaluating reading comprehension systems[EB/OL]. (2017-07-23)[2022-01-10]. https://arxiv.org/abs/1707.07328.

[13] ZHAO Z, DUA D, SINGH S. Generating natural adversarial examples[C]// Proceedings of International Conference on Learning Representations. [S. l. : s. n.], 2018.

[14] BELINKOV Y, BISK Y. Synthetic and natural noise both break neural machine translation[C]// Proceedings of International Conference on Learning Representations. [S. l. : s. n.], 2018.

[15] GAO J, LANCHANTIN J, SOFFA M L, et al. Black-box generation of adversarial text sequences to evade deep learning classifiers[C]// Proceedings of 2018 IEEE Security and Privacy Workshops. [S. l.]: IEEE, 2018: 50-56.

[16] HOSSEINI H, KANNAN S, ZHANG B, et al. Deceiving google's perspective api built for detecting toxic comments [EB/OL]. (2017-02-27) [2022-01-10]. https://arxiv.org/abs/1702.08138.

[17] PRUTHI D, DHINGRA B, LIPTON Z C. Combating adversarial misspellings with robust word recognition [C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy:[s. n.], 2019.

[18] SATO M, SUZUKI J, SHINDO H, et al. Interpretable adversarial perturbation in input embedding space for text[C]// Proceedings of the 27th International Joint Conference on Artificial Intelligence and the 23rd European Conference on Artificial Intelligence. Stockholm, Sweden:[s. n.], 2018.

[19] ZHANG H, ZHOU H, MIAO N, et al. Generating fluent adversarial examples for natural languages[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: [s. n.], 2019.

[20] SAMANTA S, MEHTA S. Towards crafting text adversarial samples[EB/OL]. (2017-07-10)[2022-01-10]. https://arxiv.org/abs/1707.02812.

[21] REN S, DENG Y, HE K, et al. Generating natural language adversarial examples through probability weighted word saliency[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy:[s. n.], 2019.

[22] WALLACE E, FENG S, KANDPAL N, et al. Universal adversarial triggers for attacking and analyzing NLP[C]// Proceedings of 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Hongkong, China:[s. n.], 2019.

[23] DONG Z, DONG Q, HAO C. HowNet and Its Computation of Meaning[C]// Proceedings of the 23rd International Conference on Computational Linguistics. Beijing, China:[s. n.], 2010.

[24] MAAS A, DALY R E, PHAM P T, et al. Learning word vectors for sentiment analysis[C]// Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Oregon, USA:[s. n.], 2011.

[25] SOCHER R, PERELYGIN A, WU J, et al. Recursive deep models for semantic compositionality over a sentiment treebank[C]//Proceedings of 2013 Conference on Empirical Methods in Natural Language Processing. Seattle, USA:[s. n.], 2013.

[26] CONNEAU A, KIELA D, SCHWENK H, et al. Supervised learning of universal sentence representations from natural language inference data[C]// Proceedings of 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark:[s. n.], 2017.

[27] DEVLIN J, CHANG M W, LEE K, et al. Bert: pre-training of deep bidirectional transformers for language understanding[EB/OL]. (2018-10-11)[2022-01-10]. https://arxiv.org/abs/1810.04805.

[28] CHEN X, SALEM A, BACKES M, et al. Badnl: Backdoor attacks against NLP models with Semantic-preserving Improvements [C]// Proceedings of ICML 2021 Workshop on Adversarial Machine Learning. [S. l. : s. n.], 2021.

[29] RADFORD A, WU J, CHILD R, et al. Language models are unsupervised multitask learners [J]. OpenAI blog, 2019, 1(8): 9.

责任编辑　董　莉