中图法分类号:TP391　　**文献标识码**:A　　**文章编号**:1006-8961(2024)07-1875-14

论文引用格式:Chen Y H, Du X, Wang D H, Wu Y, Zhu S Z and Yan Y. 2024. Universal detection method for mitigating adversarial text attacks through token loss information. Journal of Image and Graphics, 29(07):1875-1888(陈宇涵, 杜侠, 王大寒, 吴芸, 朱顺痣, 严严. 2024. 令牌损失信息的通用文本攻击检测. 中国图象图形学报, 29(07):1875-1888)[DOI:10.11834/jig.230432]

令牌损失信息的通用文本攻击检测

陈宇涵[1],杜侠[1*],王大寒[1],吴芸[1],朱顺痣[1],严严[2]

1. 厦门理工学院计算机与信息工程学院福建省模式识别与图像理解重点实验室,厦门　361024;
2. 厦门大学信息学院,厦门　361005

摘　要:**目的**　文本对抗攻击主要分为实例型攻击和通用非实例型攻击。以通用触发器(universal trigger,UniTrigger)为代表的通用非实例型攻击对文本预测任务造成严重影响,该方法通过生成特定攻击序列使得目标模型预测精度降至接近零。为了抵御通用文本触发器攻击的侵扰,本文从图像对抗性样本检测器中得到启发,提出一种基于令牌损失权重信息的对抗性文本检测方法(loss-based detect universal adversarial attack,LBD-UAA),针对UniTrigger攻击进行防御。**方法**　首先LBD-UAA分割目标样本为独立令牌序列,其次计算每个序列的令牌损失权重度量值(token-loss value,TLV)以此建立全样本序列查询表。最后基于UniTrigger攻击的扰动序列在查询表中影响值较大,将全序列查询表输入设定的差异性检测器中通过阈值阀门进行对抗性文本检测。**结果**　通过在4个数据集上进行性能检测实验,验证所提出方法的有效性。结果表明,此方法在对抗性样本识别准确率上高达97.17%,最高对抗样本召回率达到100%。与其他3种检测方法相比,LBD-UAA在真阳率和假阳率的最佳性能达到99.6%和6.8%,均实现大幅度超越。同时,通过设置先验判断将短样本检测的误判率降低约50%。**结论**　针对UniTrigger为代表的非实例通用式对抗性攻击提出LBD-UAA检测方法,并在多个数据集上取得最优的检测结果,为文本对抗检测提供一种更有效的参考机制。

关键词:文本对抗样本;通用触发器;文本分类;深度学习;对抗性检测

Universal detection method for mitigating adversarial text attacks through token loss information

Chen Yuhan[1], Du Xia[1*], Wang Dahan[1], Wu Yun[1], Zhu Shunzhi[1], Yan Yan[2]

1. School of Computer and Information Engineering, Xiamen University of Technology, Fujian Key Laboratory of Pattern Recognition and Image Understanding, Xiamen　361024, China; 2. School of Informatics, Xiamen University, Xiamen　361005, China

Abstract: **Objective** In recent years, adversarial text attacks have become a hot research problem in natural language processing security. An adversarial text attack is an malicious attack that misleads a text classifier by modifying the original text to craft an adversarial text. Natural language processing tasks, such as smishing scams (SMS), ad sales, malicious

收稿日期:2023-07-10;**修回日期**:2024-01-04;**预印本日期**:2024-01-11

*通信作者:杜侠　duxia@xmut.edu.cn

基金项目:福建省科技厅高校产学合作项目(2021H6035);厦门市留学人员科研项目(厦人社[2022]205号-02);厦门市科技计划项目(3502Z20231042):福建省自然科学基金项目(2021J011191);福厦泉国家自主创新示范项目(2022FX4)

Supported by: Industry-University Cooperation Project of Fujian Science and Technology Department (2021H6035); Xiamen Research Project for the Returned Overseas Chinese Scholars (Xiamen Human Resources Society [2022] 205-02); Xiamen Science and Technology Plan Project (3502Z20231042); Natural Science Foundation of Fujian Province, China (2021J011191); Fu-Xia-Quan National Independent Demonstration Project (2022FX4)

comments, and opinion detection, can be achieved by creating attacks corresponding to them to mislead text classifiers. A perfect text adversarial example needs to have imperceptible adversarial perturbation and unaffected syntactic-semantic correctness, which significantly increases the difficulty of the attack. The adversarial attack methods in the image domain cannot be directly applied to textual attacks due to discrete text limitation. Existing text attacks can be categorized into two dominant groups: instance-based and learning-based universal non-instance attacks. For instance-based attacks, a specific adversarial example is generated for each input. For learning-based universal non-instance attacks, universal trigger (UniTrigger) is the most representative attack, which reduces the accuracy of the objective model to near zero by generating a fixed sequence of attacks. Existing detection methods mainly tackle instance-based attacks but are seldom studied in UniTrigger attacks. Inspired by the logit-based adversarial detector in computer vision, we propose a UniTrigger defense method based on token loss weight information. **Method** For our proposed loss-based detect universal adversarial attack (LBD-UAA), we generalize the pre-training model to transform token sequences into word vector sequences to obtain the representation of token sequences in the semantic space. Then, we remove the target to compute the token positions and feed the remaining token sequence strings into the model. In this paper, we use the token loss value (TLV) metric to obtain the weight proportion of each token to build a full-sample sequence lookup table. The token sequences of non-UniTrigger attacks have less fluctuation than the adversarial examples in the variation of the TLV metric. Prior knowledge suggests that the fluctuations in the token sequence are the result of adversarial perturbations generated by UniTrigger. Hence, we envision deriving the distinct numerical differences between the TLV full-sequence lookup table and clean samples, as well as adversarial samples. Subsequently, we can employ the differential outcomes as the data representation for the samples. Building upon this approach, we can set a differential threshold to confine the magnitude of variations. If the magnitude exceeds this threshold, then the input sample will be identified as an adversarial instance. **Result** To demonstrate the efficacy of the proposed approach, we conducted performance evaluations on four widely used text classification datasets: SST-2, MR, AG, and Yelp. SST-2 and MR represent short-text datasets, while AG and Yelp encompass a variety of domain-specific news articles and website reviews, making them long-text datasets. First, we generated corresponding trigger sequences by attacking specific categories of the four text datasets through the UniTrigger attack framework. Subsequently, we blended the adversarial samples evenly with clean samples and fed them into the LBD-UAA for adversarial detection. Experimental results across the four datasets indicate that this method achieves a maximum detection rate of 97.17%, with a recall rate reaching 100%. When compared with four other detection methods, our proposed approach achieves an overall outperformance with a true positive rate of 99.6% and a false positive rate of 6.8%. Even for the challenging MR dataset, it retains a 96.2% detection rate and outperforms the state-of-the-art approaches. In the generalization experiments, we performed detection on adversarial samples generated using three attack methods from TextBugger and the PWWS attack. Results indicate that LBD-UAA achieves strong detection performance across the four different word-level attack methods, with an average true positive rate for adversarial detection reaching 86.77%, 90.98%, 90.56%, and 93.89%. This finding demonstrates that LBD-UAA possesses significant discriminative capabilities in detecting instance-specific adversarial samples, showcasing robust generalization performance. Moreover, we successfully reduced the false positive rate of short sample detection to 50% by using our proposed differential threshold setting. **Conclusion** In this paper, we follow the design idea of adversarial detection tasks in the image domain and, for the first time in the general text adversarial domain, introduce a detection method called LBD-UAA, which leverages token weight information from the perspective of token loss measurement, as measured by TLV. We are the first to detect UniTrigger attacks by using token loss weights in the adversarial text domain. This method is tailored for learning-based universal category adversarial attacks and has been evaluated for its defensive capabilities in sentiment analysis and text classification models in two short-text and long-text datasets. During the experimental process, we observed that the numerical feedback from TLV can be used to identify specific locations where perturbation sequences were added to some samples. Future work will focus on using the proposed detection method to eliminate high-risk samples, potentially allowing for the restoration of adversarial samples. We believe that LBD-UAA opens up additional possibilities for exploring future defenses against UniTrigger-type and other text-based adversarial strategies and that it can provide a more effective reference mechanism for adversarial text detection.

Key words: adversarial text examples; universal triggers; text classification; deep learning; adversarial detection

0 引 言

深度神经网络在多种机器学习任务中都取得了不俗的表现，特别在自然语言处理（natural language processing，NLP）任务中进展显著。同时，由于深度网络在结构上的复杂性和缺少可解释性，使得网络结构极易受到微小扰动的影响。对于在干净样本中添加微小且不易察觉的扰动而产生的对抗样本，模型的识别将会受到严重干扰。例如在图像中加入一些无法察觉的掩码信息（Goodfellow 等，2015），对文本进行细微的修改，或是在图像和文本上同时添加扰动的多模态攻击（Yuan 等，2023）。因此，针对对抗攻击的防御方法成为当下研究的重点。

虽然现有的对抗样本相关工作主要集中于计算机视觉领域任务中（Szegedy 等，2014；Anish 等，2018），近几年来针对 NLP 任务的对抗样本研究（Papernot 等，2016；Xu 等，2020）也逐渐出现。例如在针对垃圾邮件的恶意攻击中，攻击者可以设置特定的干扰信息以误导垃圾邮件过滤系统的识别结果。不仅如此，在现实世界中涉及到自然语言处理领域任务（Gan 和 Ng，2019；袁天昊 等，2022），如短信诈骗、广告推销、恶意评论、舆情检测、开放域问答（Fang 和 Wang，2023）等均可以通过制作与之相对应的攻击达到误导文本分类器的效果。

一个严谨的文本对抗扰动既要实现对抗扰动的隐蔽性，同时要保证语法和语义的正确性不受影响，这大大提高了文本对抗样本的制作难度。且文本的离散性特点也使得图像领域中的对抗攻击方法（Dong 等，2018；钱申诚 等，2022）无法直接运用到文本攻击中。现有的基于 NLP 任务的对抗攻击主要包括以下两种方式：1）基于实例的攻击方法，即为每一个特定的输入生成特定的对抗样本。Gao 等人（2018）提出了一种从任意字符上干扰原始文本的攻击方法，由图像领域的快速梯度符号法（fast gradient sign method，FGSM）思想迁移到文本中，通过寻找热字符干扰字符的组成结构，提出插入、删除、修改字符的方式实现攻击目的；Ebrahimi 等人（2018）和 Jin 等人（2020）在基于词嵌入空间中寻找相近语义信息且能使文本分类器分类错误的词替换攻击策略；在同义词替换攻击基础上，Ren 等人（2019）提出一种基于网络概率加权词显著性算法 PWWS（probability weighted word saliency）。Iyyer 等人（2018）通过控制句子释义网络（syntactically controlled paraphrase network，SCPN），给定一个句子和一个目标句法形式，使得每一个实例样本经过 SCPN 训练后目标句子被翻译成所需句子以生成对抗文本，该方法首先大规模反向翻译标记其中发生的自然句法转换，然后用额外的输入训练一个神经编码器解码模型并指定目标语法。2）基于触发器的通用攻击（universal trigger，UniTrigger）方法。Wallace 等人（2019）提出的通用触发器 UniTrigger 是一个不基于特定实例添加特定攻击的文本攻击方法，通用触发器基于现有深度神经网络中“学习”生成对抗序列，这是一种更为高效的攻击方法，它能够对特定类别的批次样本添加相同的扰动，从而使得文本分类器错误分类。

相比于种类丰富的图像攻击方法检测（赵俊杰等，2023），针对于现有文本攻击的检测工作相对较少。例如，Rodriguez 和 Rojas-Galeano（2018）、Pruthi 等人（2019）发现，采用破坏语法的字符或者单词的扰动方式可以轻易通过单词拼写检查检测出对抗样本。在针对同义词替换的词级文本攻击中，Alzantot 等人（2018）和 Ren 等人（2019）通过在训练过程中加入受扰动的对抗性文本的对抗训练方式提高模型的鲁棒性。

虽然当前方法对防御实例型攻击有一定效果（Zeng 等，2023），但对通用攻击的检测效果并不理想。以 UniTrigger 为代表的非实例通用攻击方法为例，该方法利用模型整体的潜在弱点进行攻击，并利用待攻击模型的梯度信息作为引导方向寻找最佳扰动，从而通过添加扰动序列以实现误导模型的目的，最终可将神经网络模型的预测精度降低到接近零的水平。特别是 UniTrigger 的攻击不针对任何特定的模型和输入，在相似模型之间具有高度的泛化性。此外，经过 UniTrigger 生成的对抗例子（Wallace 等，2019；Le 等，2020）攻击黑盒神经网络模型同样能取得较好的效果。

目前针对 UniTrigger 攻击的防御工作相对稀缺。Le 等人（2021）提出利用“蜜罐”（DARCY）在网络中设置多个陷阱门，诱导攻击令牌的生成，从而对 UniTrigger 攻击进行检测。然而，基于 DARCY 陷阱门的防御形式需要重新训练文本分类器，从而达到诱导的目的，这不适合现实场景中文本对抗检测任务。

基于此，本文提出一种利用对抗性令牌损失信息，针对通用文本攻击的检测方法，即通过样本序列化令牌的梯度来寻找损失权重信息，以此转化为令牌损失度量值（token-loss value，TLV），并通过设定检测器阈值判断对抗性样本。

Wang和He（2021）通过分析模型的逻辑回归信息（logistic，Logits）从而在图像对抗样本检测上取得显著优势。基于Logits的对抗样本检测器只在计算机视觉应用中被提出，本文从中得到启发，索引对抗样本的令牌信息并将其转为词嵌入向量表示，通过Logits作为指导信息转化为损失度量值TLV建立对抗样本全部令牌信息的损失权重全序列查询表。而对抗样本检测器分别在对抗样本的全序列查询表和干净样本的全序列查询表中进行工作。

本文主要贡献如下：1）沿用图像领域中对抗性检测任务设计思路，在通用文本对抗领域中首次从令牌权重角度出发，利用令牌损失度量信息TLV在通用文本对抗领域上实现UniTrigger攻击的检测工作。2）提出一种针对学习式通用类别对抗攻击的检测方法（loss-based detect universal adversarial attack，LBD-UAA），并在双向长短期记忆网络（bidirectional long short-term memory，BiLSTM）模型下对情感分析、文本分类的2个短数据集及2个长数据集评估了其防御性能。3）通过在短文本数据集上加入先验阈值限制，保证LBD-UAA针对对抗样本的高检测率的同时降低了其在干净样本上的误检率。

1 相关工作

1.1 文本对抗攻击

与图像领域的对抗攻击（Zhang等，2020）不同，文本的扰动输入需要在离散的词向量空间中找到对应的令牌作为扰动的添加输入（Bajaj和Vishwakarma，2023），因此基于图像的梯度方法（Goodfellow等，2015）无法运用于文本对抗样本中。对于文本对抗样本，给定一个输入的样本$\boldsymbol{x}$和一个目标模型f，一个对抗样本的组成是$\boldsymbol{x}'=\boldsymbol{x}+\Delta\boldsymbol{x}$，通过对原始输入$\boldsymbol{x}$添加一个扰动$\Delta\boldsymbol{x}$，以实现误导目标模型的目的$f(\boldsymbol{x})=y\neq y'=\arg\max f(\boldsymbol{x}')$，其中，$\Delta\boldsymbol{x}$必须限制在最低影响文本可读性的范围内。在文本对抗样本研究中，$\Delta\boldsymbol{x}$可以是干净文本中字符级别的改动、在单词层面对随机单词的同义词替换以及UniTrigger类型攻击针对模型的脆弱点在文本中添加特定的攻击令牌。表1列出了UniTrigger在BiLSTM模型下情感分类任务中针对SST-2（Stanford sentiment treebank 2）数据集在不同类别中进行通用攻击的具体例子，且将触发神经网络模型误分类的扰动序列用下划线表示，并放置在扰动样本的头部位置。

表1 UniTrigger在BiLSTM模型中攻击SST-2数据集的示例

Table 1 Example of UniTrigger attacking SST-2 dataset in BiLSTM model

原始类别	输入（下划线表示扰动）	输出结果
消极	Deeply compassionately smoothly It's somewhat clumsy and too lethargically paced—its story about a mysterious creature with psychic abilities offers a solid and some nice chills along the way Deeply compassionately smoothly Pumpkin takes an admirable look at the hypocrisy of political correctness but it does so with such an uneven tone that you never know when humor ends and tragedy begins.	积极
积极	Ransacked sorriest garbage Visually imaginative thematically instructive and thoroughly delightful it takes us on a roller-coaster ride from innocence to experience without even a hint of that typical kiddie-flick sentimentality Ransacked sorriest garbage You'll gasp appalled and laugh outraged and possibly watchinf the spectacle of a promising young lad treading desperately in a nasty sea shed an errant tear	消极

1.2 通用对抗攻击方法

UniTrigger攻击最早在图像领域被提出，Moosavi-Dezfooli等人（2017）首先提出通用攻击概念；之后Behjati等人（2019）从梯度投影的角度出发，在文本分类器中实现对抗攻击；随后Wallace等人（2019）提出在梯度引导中产生面向通用类别的文本对抗样本UniTrigger，该方法在文本分类、阅读理解和文本生成任务中均取得一定成效。在UniTrigger任务中设目标函数$F(\boldsymbol{x},\theta)$，其中用θ参数化目标神经网络。在训练数据集$\boldsymbol{D}_{\text{train}}\leftarrow\{\boldsymbol{x},\boldsymbol{y}\}_i^N$中训练目标神经网络，其中$\boldsymbol{N}$是训练样本集和。在测试数据集中提取待攻击类别标签$\boldsymbol{C}$的集合，$\boldsymbol{C}$为目标文本$\boldsymbol{x}$

的真实标签。指定文本$\boldsymbol{x}$为特定类别的概率用$\boldsymbol{L}$表示。UniTrigger生成由k个标记令牌组成的固定短语$\boldsymbol{S}$,即触发器,并在文本$\boldsymbol{x}$中的任意位置插入固定短语$\boldsymbol{S}$,用于攻击目标神经网络$F(\boldsymbol{x})$输出目标标签$\boldsymbol{L}$。由式(1)优化真实类别$\boldsymbol{C}$的攻击集合$\boldsymbol{D}_{\text{attack}}$,具体为

$$\min_{s} Loss_{\text{L}} = -\sum_{i,y_i \neq \text{L}} \log(F(\boldsymbol{S} \oplus \boldsymbol{x}_i, \theta)_{\text{L}}) \tag{1}$$

式中,⊕代表令牌连接符号。式(1)首先将S初始化定位为一个中等令牌词汇(例如:“the”,“the”,“the”),UniTrigger使用波束搜索的方法在集合$\boldsymbol{D}_{\text{attack}}$以小批量优化的方法更新触发器$\boldsymbol{S}$得到最佳的扰动令牌,从式(1)的初始化令牌中进行优化直至优化函数$Loss_{\text{L}}$收敛,并将最后一组的扰动令牌作为类别$\boldsymbol{C}$的通用触发器。

1.3 文本攻击的性能和检测

表2展示了Le等人(2021)在MR(morie review dataset)数据集和SST(stanford sontiment treebnk)数据集上针对字符替换攻击HotFlip(Ebrahimi等,2018)、词级编辑攻击TextBugger(Li等,2019)、同义词替换攻击TextFooler(Jin等,2020)以及UniTrigger进行攻击的性能测试情况。其中,TextFooler和TextBugger通过在词嵌入空间中寻找语义相似度高的单词进行替换,从而保证输入的扰动文本在语义相似度上的偏差处于可控范围之内,但其攻击性能存在较大的局限。然而,UniTrigger仅添加少量的攻击令牌就能使得卷积神经网络(convolutional neural network,CNN)模型性能大幅下降,与其他类型攻击方式相比更高效。即使运用通用句子编辑器USE(universal sentence encoder)(Cer等,2018)进行词级检测和移除,UniTrigger依然能够将CNN模型的预测准确率降到30%,明显优于其他方法。为了防御UniTrigger型攻击,本文利用UniTrigger攻击令牌在对抗性文本中词显著性的特点,通过梯度转化为损失权重信息进行对抗样本检测。

表2 Le等人(2021)在MR数据集和SST数据集上测试文本攻击的性能情况

Table 2 Performance of Le et al. (2021) in testing text attacks on pairs of MR dataset and SST dataset

/%

攻击类型	MR数据集		SST数据集	
	消极	积极	消极	积极
HotFlip	91.9	48.8	90.1	60.3
TextFooler	70.4	25.9	65.5	34.3
TextBugger	91.9	46.7	87.9	63.8
UniTrigger	1.7	0.4	2.8	0.2
USE+UniTrigger	29.2	28.3	30	28.1

2 方 法

2.1 对抗检测问题描述

所提出的防御方法属于对抗检测的范畴,主要目的为保护目标模型免受通用式文本对抗样本的攻击,同时揭示模型对于对抗样本和原始样本在深度神经网络中的差异性。

对于一个给定的n分类的文本分类器$F: \boldsymbol{X} \rightarrow \boldsymbol{Y}$,UniTrigger攻击的目标是对式(2)中的损失L进行优化,从而得到该批次信息中的最优解,具体为

$$\arg\min_{t_{\text{adv}}} \left[L(\tilde{y}, F(\boldsymbol{t}_{\text{adv}}; \boldsymbol{t})) \right] \tag{2}$$

式中,$\boldsymbol{t}$为原始文本,$\boldsymbol{t}_{\text{adv}}$为添加在文本中的扰动序列,优化式(2)使得对于目标类$\tilde{y}$的损失最小。

定义一个被UniTrigger成功攻击的样本$\boldsymbol{X}' = [\boldsymbol{x}_1', \boldsymbol{x}_2', \cdots, \boldsymbol{x}_n']$,一个没有被污染的干净样本$\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n]$。其中$\boldsymbol{X}' \in \boldsymbol{C}$为指定类别文本$\boldsymbol{C}$中的数据分布,$\boldsymbol{x}_1' \sim \boldsymbol{x}_n'$代表对抗文本中由字符或者词串组成的令牌。$\boldsymbol{X}'$中含有长度为$k$的对抗序列$\boldsymbol{t}_{\text{adv}} = \{\boldsymbol{t}_1, \boldsymbol{t}_2, \cdots, \boldsymbol{t}_k\}$插入到文本样本$\boldsymbol{x}_1' \sim \boldsymbol{x}_n'$的某一位置,所提出的LBD-UAA目的是区分$\boldsymbol{X}'$和$\boldsymbol{X}$,干净样本和对抗样本均匀分布在批次样本中,文本分类器$F: \boldsymbol{X} \rightarrow \boldsymbol{Y}$对当前输入样本的令牌序列进行预处理,从$\boldsymbol{X}' = [\boldsymbol{x}_1', \boldsymbol{x}_2', \cdots, \boldsymbol{x}_n']$中分割出样本的令牌序列,通过$\boldsymbol{t}_i$在分类器$F$中损失权重占比信息辨别样本的性质,即LBD-UAA的目的可以表示为

$$\text{Detector}(\text{argmax}\, TLV(\boldsymbol{t}_{\text{adv}}, \boldsymbol{X})) \sim \text{adversarial} \tag{3}$$

图1显示了LBD-UAA方法的整体流程,分为TLV计算层和对抗性检测层,其细节将在2.2节及2.3节中详细介绍,包括对抗样本的令牌损失度量信息计算、TLV全序列查询表的建立和对抗样本检测器的方法。

2.2 令牌损失信息计算

Mosca等人(2021)在基于语义相似度的最小词替换语义对抗攻击中,通过找到目标函数模型对于

被攻击词在深度神经网络中的敏感程度以发现对抗性，本文同样从被UniTrigger攻击的样本中识别$f(\boldsymbol{x}')$与$f(\boldsymbol{x})$在模型中的差异性。图1详细展示了所提方法的整体流程。首先将对抗样本中干净样本均匀混合成待检测样本集$\boldsymbol{N}$，从$\boldsymbol{N}$中随机取样$\boldsymbol{X}$，与图像对抗样本不同，文本样本本身具有强离散性，且UniTrigger是通过添加扰动令牌的攻击方式，因此将文本集合$\boldsymbol{X}$拆分成$\boldsymbol{X}=[\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n]$，形成独立的令牌序列块。并通过令牌损失度量TLV评估模型$F(\boldsymbol{X})$对给定输入的反应。具体的方法是对令牌序列$\boldsymbol{x}$通过预训练模型利用式(4)转为300维词向量$\boldsymbol{V}(\boldsymbol{x})=\left[\boldsymbol{V}(\boldsymbol{x}_1),\boldsymbol{V}(\boldsymbol{x}_2),\cdots,\boldsymbol{V}(\boldsymbol{x}_n)\right]$，具体为

$$\boldsymbol{V}(\boldsymbol{x}_k)=\frac{1}{2C}\times\sum\boldsymbol{V}(\boldsymbol{x}_{k-c})+\sum\boldsymbol{V}(\boldsymbol{x}_{k+c}) \tag{4}$$

式中，C代表词向量大小。对于所有令牌所组成的$\boldsymbol{V}(\boldsymbol{x}_i)$，集成为待检测令牌序列集$\boldsymbol{V}(\boldsymbol{x}_i)\in\boldsymbol{D}$，假定检测第$k$个令牌序列，则在序列集$\boldsymbol{D}$中抽取第$k$个$\boldsymbol{V}(\boldsymbol{x}_k)$，使得$D_{\mathrm{dec}}=\sum_{i=1}^{k-1}\boldsymbol{v}(\boldsymbol{x}_i)+\sum_{j=k+1}^{n}\boldsymbol{v}(\boldsymbol{x}_j)$，将$\boldsymbol{D}_{\mathrm{dec}}$输入到文本分类器$F(X)$，用$Log(\boldsymbol{x})$函数代表目标参数在模型中的Logits值，得到梯度信息$Log(\boldsymbol{x}_k)$的计算式，具体为

$$\nabla Log(\boldsymbol{x}_k)=\nabla Log\left(F\left(\boldsymbol{D}_{\mathrm{dec}}\right)\right) \tag{5}$$

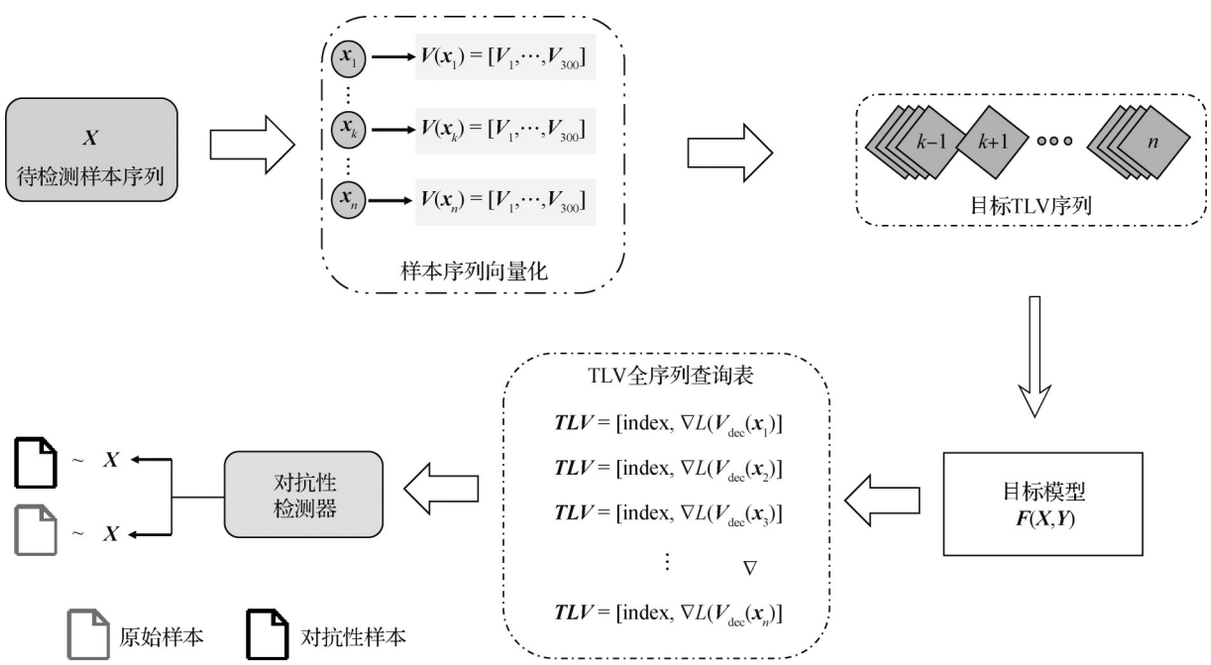


图1　LBD-UAA方法的概述

Fig. 1　Overview of the LBD-UAA method

2.3　对抗攻击检测算法

在检测对抗样本的过程中发现，受到干扰的全序列查询表与干净样本查询表相比存在较大的差异，这是因为受到干扰的样本在令牌损失中的占比会更大，在UniTrigger攻击序列中具体表现为对应损失度量值TLV远小于正常令牌。

对于输入样本$\boldsymbol{X}$，第i个令牌损失权重信息$\boldsymbol{TLV}$可表示为

$$\boldsymbol{TLV}=\left[\mathrm{index},\nabla L_{\mathrm{CE}}\left(V_{\mathrm{dec}}(\boldsymbol{x}_i)\right)\right] \tag{6}$$

式中，$\boldsymbol{L}_{\mathrm{CE}}$表示交叉熵损失函数，index表示在全令牌序列查询表中的标记位。如表3所示，对输入样本中的每个令牌序列计算$\boldsymbol{TLV}$组成一张全序列查询表，该表显示了对应的令牌在文本序列中的位置以及TLV值，其中加粗部分为示例样本表中该令牌损失占比最高位置，从先验知识可以获知该令牌为UniTrigger加入的对抗性令牌。

LBD-UAA算法的具体步骤如下：

算法1　LBD-UAA

输入：数据分布$\boldsymbol{X}$；文本分类器$\boldsymbol{F}(\boldsymbol{X})$；嵌入向量空间$\boldsymbol{V}(\boldsymbol{x})$；分类器Logits值$Log(x_i)$；损失权重信息$\boldsymbol{TLV}$；全序列查询表$\boldsymbol{T}$。

输出：$\{\boldsymbol{x}|\boldsymbol{x}\in\boldsymbol{adv}\}$；$\{\boldsymbol{x}|\boldsymbol{x}\in\boldsymbol{ori}\}$//$\boldsymbol{adv}$为对抗性

表 3　输入样本转为 TLV 度量值建立全序列查询表

Table 3　Input samples are converted to TLV metrics to build a full sequence lookup table

独立令牌	下标	***TLV*** (index, value)
ransacked	0	**1.857 661 7**
garbage	2	3.489 932 0
sorriest	1	3.681 084 6
Lovely	14	4.986 374 0
⋮	⋮	⋮
Accorsi	4	7.534 617 4
by	6	7.054 067 0

注：加粗字体表示损失占比最高值。

样本；***ori***为原始样本。

1）$\boldsymbol{X}=[\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \cdots, \boldsymbol{x}_n]$//$n$个令牌序列；

2）$\boldsymbol{F}(\boldsymbol{X})$；

3）for i in $\boldsymbol{x}$//取第k个令牌为例；

4）　$\boldsymbol{V}_{\text{dec}}=[\boldsymbol{V}(\boldsymbol{x}_1), \boldsymbol{V}(\boldsymbol{x}_2), \cdots, \boldsymbol{V}(\boldsymbol{x}_{k-1}), \boldsymbol{V}(\boldsymbol{x}_{k+1}), \cdots, \boldsymbol{V}(\boldsymbol{x}_n)]$；

5）　$\boldsymbol{V}_{\text{dec}} \rightarrow \boldsymbol{F}(\boldsymbol{X})$；

6）　$\nabla Log(\boldsymbol{x}_k)=\nabla Log\left(\boldsymbol{F}\left(\boldsymbol{D}_{\text{dec}}\right)\right)$；

7）　$\boldsymbol{TLV}=\left[k, \nabla L_{\text{CE}}\left(\boldsymbol{V}_{\text{dec}}(\boldsymbol{x}_i)\right)\right]$；

8）　end for；

9）　$\forall \boldsymbol{TLV} \in \boldsymbol{T}$//建立全序列查询表$\boldsymbol{T}$；

10）　$df=\text{argmax}(\boldsymbol{T})-\text{argmin}(\boldsymbol{T})$；

11）if $df>\alpha$：

12）　$\boldsymbol{x} \rightarrow \{\boldsymbol{x} \mid \boldsymbol{x} \in \boldsymbol{adv}\}$//对抗性样本；

13）else：

14）　$\boldsymbol{x} \rightarrow \{\boldsymbol{x} \mid \boldsymbol{x} \in \boldsymbol{ori}\}$//原始样本。

首先，将全序列查询表输入到检测器中，根据***TLV***函数值找到最大索引值和最小索引值，其中df表示查询列表中的最大差异值，上述具体表现为$df=\text{argmax}(\boldsymbol{T})-\text{argmin}(\boldsymbol{T})$，其中$\boldsymbol{T}$为查询表集合。同时，实验中发现非 UniTrigger 攻击的令牌序列在***TLV***的度量值中变化波动较小，因此将df作为输入样本在对抗性检测器的数据表征，并设置幅度变化阈值α限制。最后，若超过该阈值则判断输入样本为对抗样本，否则为干净样本。在第 3 节的实验中将详细说明对比不同类型数据集下干净样本和对抗样本在阈值α变化的表现情况。

3　实验结果及分析

3.1　实验设置

3.1.1　数据集

数据集采用在自然语言处理领域常用的 4 个公开英文数据集作为样本来源，如表 4 所示，SST-2（Wang 等，2018）是一个二元情感分析数据集，由美国斯坦福大学发布用于训练和评估情感分析模型，数据类别由正向情感和负向情感组成。MR（Pang 和 Lee，2005）源于 IMDB（internet movie database ）网站的电影评论，其情感标签分为两类，用于表达对电影的评价，上述两种数据集均属于短文本数据集。长文本数据集则选用 AG（ag news topic classification dataset）（Zhang 等，2015）和 Yelp（Zhang 等，2015），其中，AG 数据集包含 4 个不同领域的新闻文章（Business，Sports，Science/ Technology，World），每篇文章都分配了一个类别标签，用于表示所属的领域类别。Yelp 由 Yelp 网站上的用户评论组成，表示用户对商家的评价用于两个类别的情感分类判断，该数据集的平均样本长度最长为 581.56 个字符数。所有实验取每个数据集的测试集作为实验样品，均匀混合对抗样本和原始样本。

3.1.2　分类模型

在模型场景设置中，选取文本分类任务中最为常见的 BiLSTM 模型（McCann 等，2017）作为防御对象，同时在 UniTrigger 攻击下作为受害者模型用于保持模型样本测试的一致性。如表 4 所示，实验选用上述 4 个数据集分别训练文本分类任务，作为生成对抗样本的载体。从测试结果中可以看出，除 MR 数据集外，其余数据集在 BiLSTM 模型下训练产生的分类器能够保持较高的准确率。具体来说，文本分类器$f(\boldsymbol{x})$准确率分别可达 97.08%、93.20%、87.17% 和 73.52%。此外，在进一步的实验中发现 LBD-UAA 在应对分类模型精度不高的情况下仍然具有较高的对抗样本检测率的优势，即 LBD-UAA 是一种高性能的检测方法。

3.2　对抗样本生成

采用 UniTrigger 论文中提供的代码框架作为基线以生成对抗性样本。由于 4 个数据集均为英文数据集，因此实验采用 300 维的 word2vec 词嵌入向量（Li 等，2018）对所有输入的样本生成对应的词向量空

间表示。所有实验均在训练集中搜索扰动序列，并于测试集中评估效果和生成对应类别的对抗样本。

按照原攻击定义通过重复3次“the”初始化扰动序列。每个批次的大小为32，并循环迭代计算当前批次中最优扰动，直到连续5轮最优扰动不变，且整体迭代次数超过30批次以确定对应攻击序列。在所有测试集中添加扰动序列并经过模型分类选取成功误导模型分类出错的样本作为对抗样本；在所有未被攻击且分类正确的测试集上选取样本作为干净样本。表5为在4个数据集中对所有类别进行攻击并产生的扰动序列以及对应类别在受到UniTrigger攻击后模型的分类精度，UAA-acc指标数据表示添加对应扰动序列后模型分类的准确率。从该指标中可以看出，该攻击方法具有较强的普遍攻击性，特别是在AG下的Sports类攻击中达到1.57%的攻击效果。

表4　实验数据集

Table 4　Dataset statistics

数据集	任务	语言	类别数	训练集规模	测试集规模	文本平均长度	准确率(BiLSTM)/%
SST-2	情感分析	英语	2	6 920	1 821	19.28	87.17
AG	新闻分类	英语	4	120 000	7 600	43.64	93.20
MR	情感分析	英语	2	17 976	1 067	20.40	73.52
Yelp	情感分析	英语	2	560 000	38 000	581.56	97.08

表5　基于UniTrigger攻击下各类别扰动序列

Table 5　Perturbation sequences for each category under UniTrigger-based attacks

数据集	类别	扰动序列	UAA-acc/%
SST-2	Positive	Ransacked sorriest garbage	14.49
	Nagative	deeply compassionately smoothly	8.18
MR	Positive	Buttercup actuaci flavorless	7.36
	Nagative	Scrumptious scrumptious sparklingly	14.33
Yelp	Positive	STALE LeKisha people.Tons sigothers	2.38
	Nagative	Creamiest abodigas OPI-celebrities.Chances	29.67
AG	Business	AIVD hominoids okapi	2.00
	Sports	1.5B. CRC32 Firaxis	1.57
	Science	JOINS sharemarket Yukiko	26.05
	World	Colts.com KTAR NFLPA	25.84

3.3　检测性能

3.3.1　定性实验

文本在所有实验中随机选取数量相同的对抗样本与原始样本均匀混合形成待检测样本集，分别在4个数据集对应类别样本集上测试LBD-UAA检测方法的效果，并在实验中设置以下3个评价指标。

1）Precision。用于表示LBD-UAA在样本集上判定为对抗样本的样本中真实为对抗样本的比例，是衡量对抗性样本检测常用的指标。

2）Recall。用于表示LBD-UAA在样本集上判定为对抗样本的样本在所有对抗样本中的比例，是衡量对抗性样本检测常用的指标。

3）F1。综合Precision和Recall的平均评价指标，是衡量对抗性样本检测常用的指标。

表6展现了LBD-UAA策略下对所有样本的检测效果。可以看出，当前方法在4个数据集上均达到较高的检测性能，绝大部分样本分类精度能达到95%±2.0%。AG数据集中Business类别文本和Yelp数据集中Positive类别文本，其对抗样本检测精度高达97.17%和96.98%，而在SST-2数据集中Negative

类别则只有73.86%。分析原因主要与长数据样本在受到UniTrigger攻击时所添加的3个令牌序列有关，对比Yelp数据集中样本平均长度为581.56而言，3个令牌序列的扰动在损失度量信息***TLV***上占比高，即对抗样本中对抗性令牌与原始样本令牌的差异性更大。

召回率(Recall)是衡量样本被识别为对抗样本的指标，LBD-UAA在召回率上的性能普遍要优于准确率，即对于对抗样本的识别能力高。虽然小部分干净样本也被误判为对抗样本，但在现实世界中检测到对抗样本的需求远比一个干净样本被误判为对抗样本更为重要。

表6　LBD-UAA检测4个数据集对抗样本的实验结果

Table 6　Experimental results of LBD-UAA detection of four datasets against samples

/%

数据集	目标类别	Precision	Recall	F1
SST-2	Positive	**86.55**	**98.55**	**92.16**
	Negative	73.86	96.15	83.55
MR	Positive	89.17	73.43	80.52
	Negative	**96.14**	**98.59**	**97.35**
Yelp	Positive	**96.98**	**100.00**	**98.47**
	Negative	95.46	99.32	97.35
AG	Business	**97.17**	95.85	96.50
	Sports	96.74	**100.00**	**98.36**
	Science	92.92	92.92	92.92
	World	96.44	**100.00**	98.19

注：加粗字体表示每类数据集下各指标最优结果。

3.3.2　定量实验

本文所提出的LBD-UAA对抗性文本检测方法，将待测样本令牌序列化并通过式(6)获取文本分类器$f(x)$对每个独立令牌的敏感程度，进而建立文本全序列查询表，根据***TLV***度量值数据分布差异性进行对抗性检测。如算法1中所示，差异性检测器α的变化影响最终的判定性能，在4个数据集的不同类型文本载体上进行了阈值判定实验，确定阈值设定幅度在0.2～1.6之间具有判别意义。图2和表7中给出短序列文本数据集SST-2和长序列文本AG数据集上关于阈值α变化的定量实验。F1具有综合衡量准确率和召回率的特点，因此选其作为图2评价指标。从图2可以看出，F1数值随着α升高而上升，并在1.2时达到最高点。分析认为这是因为干净样本中文本长度不一致且训练过程中存在某个令牌序列在神经网络中赋予权重值过大，导致***TLV***占比较高，从而可能被错误判断为对抗性样本，因此随着阈值上升，样本误判率下降、F1数值上升。UniTrigger攻击针对分类器在样本语义空间中的脆弱位置添加恶意扰动，实现误导分类器的目的。UniTrigger攻击下存在极小部分恶意扰动转化为***TLV***度量后与正常令牌序列差异性较小，这种攻击方式更加隐蔽且刚好使得恶意扰动位于分类界面临界状态位置，从而出现随着α继续上升小部分对抗性样本逃过检测，导致F1出现下降的现象。因此，从以上F1数值波动实验结果中，最终选取$\alpha=1.2$作为差异检测器阈值。

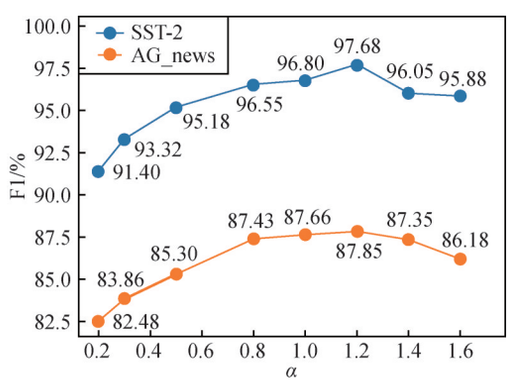


图2　阈值α对于LBD-UAA性能的影响

Fig. 2　Effect of threshold α on the performance of LBD-UAA

3.3.3　对比实验

实验将LBD-UAA与以下4种对抗样本检测算法进行比较：

1)OOD(out of distribution)(Smith和Gal，2018)。假设对抗样本在语义空间中距离训练模型的数据分布具有较远的距离，且具有高不确定性，即用高熵判断对抗样本。

2) LID (local intrinsic dimensionality) (Ma等，2018)。通过LID在对抗性区域进行描述，将对抗区域作为一个特征用于检测对抗性样本。

3)USE(Cer等，2018)。通过对输入样本的语义分数判断是否为对抗性样本。

4)DARCY(Le等，2021)。借助网络安全中“蜜罐”的方法，在网络中设置扰动陷阱，诱导特定的扰动生成，从而判断对抗性样本。

表7　LBD-UAA在SST-2和AG数据集阈值定量实验结果

Table 7　Experimental results of LBD-UAA threshold quantification in SST-2 and AG datasets

/%

阈值	AG			SST-2		
	Precision	Recall	F1	Precision	Recall	F1
1.6	96.76	95.10	95.88	82.28	90.52	86.18
1.4	96.31	95.59	96.05	81.32	95.00	87.35
1.2	95.82	97.19	**97.68**	80.20	97.35	**87.85**
1.0	94.90	98.79	96.80	79.72	99.12	87.66
0.8	93.63	99.67	96.55	78.14	99.56	87.43
0.5	90.89	100.00	95.18	74.53	100.00	85.30
0.3	87.62	100.00	93.32	72.41	100.00	83.86
0.2	84.42	100.00	91.40	70.35	100.00	82.48

注：加粗字体表示每类数据集下F1综合指标最优结果。

表8展示了LBD-UAA与其他4种检测方法在UniTrigger攻击下的性能比较，其中，真阳率（true positive rate，TPR）表示模型正确识别出对抗样本的比率，假阳率（false positive rate，FPR）表示将原始样本识别为对抗样本的比率。实验期望取得高TPR低FPR的效果。从实验结果得知，LBD-UAA在对抗性样本识别率TPR上均具有最优水平，虽然在SST-2和MR数据集上FPR性能略低于DARCY，但不同于DARCY需要对分类器重新训练以实现注入陷阱，所提方法无需重新训练网络的对抗性检测能力，因而更符合现实检测任务。其综合性能明显优于其他4种检测机制，在AG数据集上TPR达到99.6%，而FPR只有6.8%。此外，结合表4数据，虽然原始分类器在MR数据集上分类准确率并不高，但在MR对抗样本检测中仍能保有96.2%的检测率，说明所提检测方法在不同精度的分类器模型中均能保有较高的对抗样本检测率。

3.4　泛化实验

在AG数据集4个文本类别上针对实例型词级对抗性攻击方法TextBugger和PWWS进行泛化性检测。TextBugger通过对文本进行随机空格插入，删除部分字符，以及对视觉形状上相似的字符进行替换，例如“O”改为数字0，“a”替换为@以及对“L”或“l”替换为数字1等，在不影响文本可读性条件下进行对抗性干扰。PWWS通过贪婪算法对文本进行同义词替换产生对抗性攻击。

实验在TextBugger 3种攻击方法和PWWS在AG数据集中产生的对抗性样本进行检测。如表9所

表8　在UniTrigger攻击下所有目标类别的平均对抗检测性能

Table 8　Average adversarial detection performance for all target classes under UniTrigger attack

/%

方法	SST-2数据集		MR数据集		AG数据集	
	TPR	FPR	TPR	FPR	TPR	FPR
OOD（Smith和Gal，2018）	51.8	47.3	51.0	49.0	47.7	51.5
USE（Cer等，2018）	68.7	51.3	77.7	46.1	86.9	29.6
LID（Ma等，2018）	41.3	41.3	50.6	44.1	56.3	45.3
DARCY（Le等，2021）	89.4	**3.2**	87.6	**2.9**	**99.8**	9.3
LBD-UAA（本文）	**99.5**	28.6	**96.2**	16.8	99.6	**6.8**

注：加粗字体表示各数据集中的最优结果。

示，LBD-UAA 在 4 种不同的词级攻击方法上均取得较高的检测性能，平均 TPR 对抗性检测率达至 86.77%，90.98%，90.56% 和 93.89%。这说明 LBD-UAA 在实例型对抗性样本的检测任务中同样兼具较高的辨别能力，泛化性能强。

表 9　LBD-UAA 在 TextBugger 和 PWWS 下对抗性检测性能

Table 9　Adversarial detection performance of LBD-UAA under TextBugger and PWWS

/%

攻击方法	目标类别	TPR	FPR
TextBugger_Insert	World	97.04	3.54
	Sports	96.90	6.19
	Business	67.36	5.26
	Science	85.79	8.57
	平均值	86.77	5.89
TextBugger_Delete	World	97.47	5.05
	Sports	96.75	6.13
	Business	81.25	5.00
	Science	88.48	8.22
	平均值	90.98	6.10
TextBugger_Replace	World	94.84	3.91
	Sports	95.46	5.38
	Business	87.71	5.26
	Science	84.26	8.16
	平均值	90.56	5.67
PWWS	World	97.63	3.65
	Sports	98.07	7.69
	Business	91.96	3.75
	Science	87.94	8.19
	平均值	93.89	5.82

3.5　消融实验

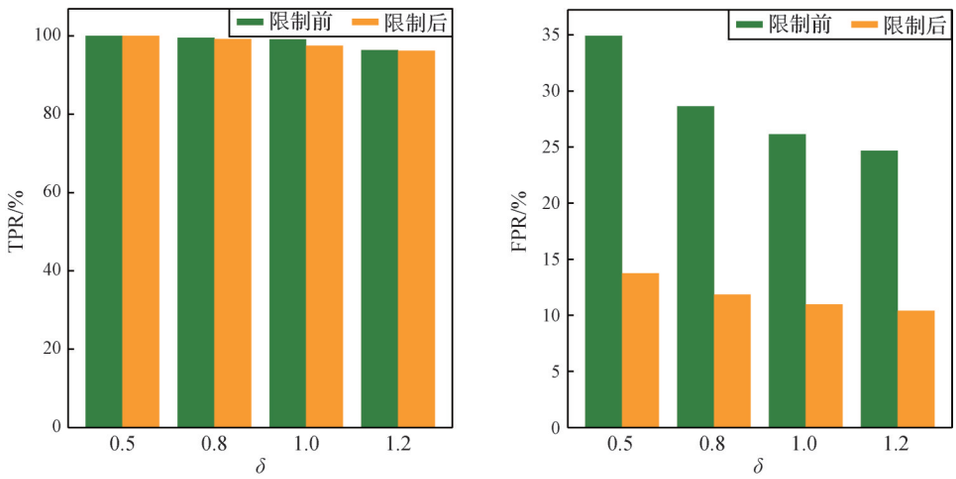
在实验中发现 2 个短文本数据集上的 FPR 值较高，这说明在短文本中误判率较高。通过分析认为这是由于短原始样本在分类器训练过程中，有限的令牌序列使得对于部分令牌在词向量空间中的权重占比高，因此这些没有被分类器很好拟合语义特征的样本被检测器误判。基于上述发现，本文在短数据样本中加入序列令牌 TLV 极小值限制，在多次实验后选取对应数据集中所有批次样本中令牌序列 TLV 极小值 x_i 求均值，即 $\delta = \frac{\sum_{i=1}^{n} TLV(x_i)}{n}$ 加入到检测器中作为判断对抗性样本的辅助信息。如图 3 所示，添加 δ 限制后 FPR 误判率明显下降，在 MR 数据集和 SST-2 数据集中每组数据平均下降约 50%，且对应的 TPR 并没有发生显著的下降，因而证明在短文本数据集中加入 TLV 极小值的先验阈值限制可以有效降低短序列样本被误判的概率。

4　结　论

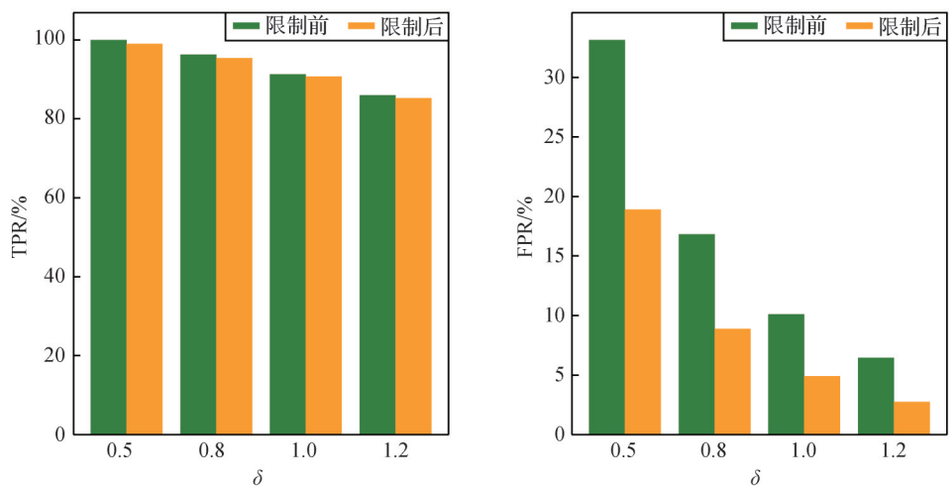
对抗性文本攻击的存在是 NLP 模型在高风险应用场景中部署的主要障碍，然而随着近几年深度学习的迅速发展，关于对抗性文本攻击的防御方法逐渐获得关注。而以 UniTrigger 为代表的非实例式通用对抗攻击相较之下要比实例对抗攻击的受众范围更广且攻击效率更高，对于攻击者而言也更具有现实意义。因此，针对 UniTrigger 为代表的非实例通用式对抗性攻击提出 LBD-UAA 检测方法。LBD-UAA 是利用对抗性文本的令牌序列输入模型并转化为损失权重信息，提出用 ***TLV*** 度量令牌序列使之以数值化形式表现并建立全序列查询表遍历输入待测样本的全部令牌信息，并在查询表中发现对抗性扰动与原始令牌序列相比具有明显的差异性。因此，通过调节检测器 α 的阈值大小进行对抗样本检测。

此外，在 2 个短样本数据集和 2 个长样本数据集上对 LBD-UAA 的性能进行检测。从检测结果可以看出，其在每个类别数据集中均取得较好的性能，Precison 指标最高达到 97.17% 且对于对抗样本的召回率达到 100%。LBD-UAA 与其他 4 种方法相比，在 TPR 和 FPR 性能上均能达到极大的提升（99.6%，6.8%），并且在针对短序列样本语义空间分布不均衡导致 FPR 值较高的情况下，引入先验判断阈值限制，使得短样本 FPR 下降幅度高达 50%。

同时，在实验过程中发现通过 ***TLV*** 的数值反馈能够查询到部分样本具体的扰动序列添加位置。在未来的工作中将尝试利用该特性检测高风险性样本，定位可能扰动位置，还原对抗性样本的工作。综



(a) SST-2数据集



(b) MR数据集

图 3　在SST-2和MR数据集上对TLV极小值的限制

Fig. 3　Constraints on TLV minima on SST-2 and MR datasets ((a) SST-2 dataset; (b) MR dataset)

上所述，LBD-UAA可以为UniTrigger类型及其他文本对抗防御策略提供更多探索的可能性。

参考文献(References)

Anish A, Carlini N and Wagner D. 2018. Obfuscated gradients give a false sense of security: circumventing defenses to adversarial examples//Proceedings of the 35th International Conference on Machine Learning (ICML 2018). Stockholm, Sweden: PMLR: 274-283

Alzantot M, Sharma Y, Elgohary A, Ho B J, Srivastava M B and Chang K W. 2018. Generating natural language adversarial examples//Proceeding of the Empirical Methods in Natural Language Processing (EMNLP 2018). Brussels, Belgium: Association for Computational Linguistics: 2890-2896 [DOI: 10.18653/v1/d18-1316]

Behjati M, Moosavi-Dezfooli S M, Baghshah M S and Frossard P. 2019. Universal adversarial attacks on text classifiers//Proceedings of 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton, UK: IEEE: 7345-7349 [DOI: 10.1109/ICASSP.2019.8682430]

Bajaj A and Vishwakarma D K. 2023. Evading text based emotion detection mechanism via adversarial attacks. Neurocomputing, 558: #126787 [DOI: 10.1016/J.NEUCOM.2023.126787]

Cer D, Yang Y F, Kong S Y, Hua N, Limtiaco N, John R S, Constant N, Guajardo-Cespedes M, Yuan S, Tar C, Strope B and Kurzweil R. 2018. Universal sentence encoder for English//Proceedings of 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. Brussels, Belgium: Association for Computational Linguistics: 169-174 [DOI: 10.18653/v1/d18-2029]

Dong Y P, Liao F Z, Pang T Y, Su H, Zhu J, Hu X L and Li J G. 2018. Boosting adversarial attacks with momentum//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2018). Salt Lake City, USA: IEEE: 9185-9193 [DOI: 10.1109/cvpr.2018.00957]

Ebrahimi J, Rao A Y, Lowd D and Dou D J. 2018. HotFlip: white-box adversarial examples for text classification//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2019). Melbourne, Australia: Association for Computational Linguistics: 31-36 [DOI: 10.18653/v1/P18-2006]

Fang X J and Wang W. 2023. Defending machine reading comprehension against question-targeted attacks//Proceedings of 2023 International Joint Conference on Neural Networks (IJCNN). Gold Coast, Australia: IEEE: 1-8 [DOI: 10.1109/IJCNN54540.2023.10191697]

Mosca E, Wich M and Groh G. 2021. Understanding and interpreting the impact of user context in hate speech detection//Proceedings of the 9th International Workshop on Natural Language Processing for Social Media. [s.l.]: Association for Computational Linguistics: 91-102 [DOI: 10.18653/v1/2021.socialnlp-1.8]

Goodfellow I, Shlens J and Szegedy C. 2015. Explaining and harnessing adversarial examples//Proceedings of 2015 International Conference on Learning Representations (ICLR 2015)

Gao J, Lanchantin J, Soffa M L and Qi Y J. 2018. Black-box generation of adversarial text sequences to evade deep learning classifiers//Proceedings of 2018 IEEE Security and Privacy Workshops. San Francisco, USA: IEEE: 50-56 [DOI: 10.1109/SPW.2018.00016]

Iyyer M, Wieting J, Gimpel K and Zettlemoyer L. 2018. Adversarial example generation with syntactically controlled paraphrase networks//Proceedings of 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. New Orleans, USA: Association for Computational Linguistics: 1875-1885 [DOI: 10.18653/v1/n18-1170]

Jin D, Jin Z J, Zhou J T and Szolovits P. 2020. Is BERT really robust? a strong baseline for natural language attack on text classification and entailment//Proceedings of 2020 Association for the Advancement of Artificial Intelligence (AAAI 2020). New York, USA: AAAI: 8018-8025 [DOI: 10.1609/aaai.v34i05.6311]

Yuan L F, Zhang Y C, Chen Y Y and Wei W. 2023. Bridge the gap between CV and NLP! A gradient-based textual adversarial attack framework//Proceedings of 2023 Association for Computational Linguistics (ACL). Toronto, Canada: Association for Computational Linguistics: 7132-7146 [DOI: 10.18653/V1/2023.finding-acl.446]

Le T, Park N and Lee D. 2021. A sweet rabbit hole by DARCY: using honeypots to detect universal trigger's adversarial attacks//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. [s.l.]: Association for Computational Linguistics: 3831-3844 [DOI: 10.18653/v1/2021.acl-lOng.296]

Le T, Wang S H and Lee D. 2020. MALCOM: generating malicious comments to attack neural fake news detection models//Proceedings of 2020 International Conference on Data Mining (ICDM 2020). Sorrento, Italy: IEEE: 282-291 [DOI: 10.1109/ICDM50108.2020.00037]

Li J F, Ji S L, Du T Y, Li B and Wang T. 2019. TextBugger: generating adversarial text against real-world applications//Proceedings of 2019 Network and Distributed System Security Symposium (NDSS 2019). San Diego, USA: The Internet Society [DOI: 10.14722/ndss.2019.23138]

Li S, Zhao Z, Hu R F, Li W S, Liu T and Du X Y. 2018. Analogical reasoning on Chinese morphological and semantic relations//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018). Melbourne, Australia: Association for Computational Linguistics: 138-143 [DOI: 10.18653/v1/p18-2023]

McCann B, Bradbury J, Xiong C M and Socher R. 2017. Learned in translation: contextualized word vectors//Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS). Long Beach, USA: Curran Associates Inc.: 6297-6308 [DOI: 10.5555/3295222.3295377]

Ma X J, Li B, Wang Y S, Erfani S M, Wijewickrema S N R, Schoenebeck G, Song D, Houle M E and Bailey J. 2018. Characterizing adversarial subspaces using local intrinsic dimensionality//Proceedings of the 6th International Conference on Learning Representations (ICLR 2018). Vancouver, Canada: ICLR

Moosavi-Dezfooli S M, Fawzi A, Fawzi O and Frossard P. 2017. Universal adversarial perturbations//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017). Honolulu, USA: IEEE: 1765-1773 [DOI: 10.1109/CVPR.2017.17]

Pang B and Lee L. 2005. Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales//Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. Ann Arbor, Michigan, USA: Association for Computational Linguistics: 115-124 [DOI: 10.3115/1219840.1219855]

Gan W C and Ng H T. 2019. Improving the robustness of question answering systems to question paraphrasing//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019). Florence, Italy: Association for Computational Linguistics: 6065-6075 [DOI: 10.18653/v1/p19-1610]

Pruthi D, Dhingra B and Lipton Z C. 2019. Combating adversarial misspellings with robust word recognition//Proceedings of the 57th Association for Computational Linguistics (ACL 2019). Florence, Italy: Association for Computational Linguistics: 5582-5591 [DOI: 10.18653/v1/p19-1561]

Papernot N, McDaniel P, Swami A and Harang R. 2016. Crafting adversarial input sequences for recurrent neural networks//Proceedings of 2016 Military Communications Conference (MILCOM 2016). Baltimore, USA: IEEE: 49-54 [DOI: 10.1109/MILCOM.2016.7795300]

Qian S C, Wen Y H, Ma Y F and Mao X W. 2022. Adversial sample attack and defense methods based on deep neural networks. Information Security and Technology, 13(5): 77-86 (钱申诚, 文宇恒, 马耀飞, 毛鑫唯. 2022. 基于深度神经网络的对抗样本攻击

与防御方法研究. 网络空间安全, 13(5): 77-86)[DOI: 10.3969/j.issn.1674-9456.2022.05.014]

Rodriguez N and Rojas-Galeano S. 2018. Shielding Google's language toxicity model against adversarial attacks [EB/OL]. [2023-05-29]. https://arxiv.org/pdf/1801.01828.pdf

Ren S H, Deng Y H, He K and Che W X. 2019. Generating natural language adversarial examples through probability weighted word saliency//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019). Florence, Italy: Association for Computational Linguistics: 1085-1097 [DOI: 10.18653/v1/p19-1103]

Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I J and Fergus R. 2014. Intriguing properties of neural networks//Proceedings of the 2nd International Conference on Learning Representations (ICLR 2014). Banff, Canada: ICLR

Smith L and Gal Y. 2018. Understanding measures of uncertainty for adversarial example detection//Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence (UAI 2019). Monterey, USA: AUAI: 560-569

Wallace E, Feng S, Kandpal N, Gardner M and Singh S. 2019. Universal adversarial triggers for attacking and analyzing NLP//Proceedings of 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: Association for Computational Linguistics: 2153-2163 [DOI: 10.18653/v1/D19-1221]

Wang A, Singh A, Michael J, Hill F, Levy O and Bowman S R. 2018. GLUE: a multi-task benchmark and analysis platform for natural language understanding//Proceedings of 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. Brussels, Belgium: Association for Computational Linguistics: 353-355 [DOI: 10.18653/v1/w18-5446]

Wang X S and He K. 2021. Enhancing the transferability of adversarial attacks through variance tuning//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2021). Nashville, USA: IEEE: 1924-1933 [DOI: 10.1109/CVPR46437.2021.00196]

Xu H, Ma Y, Liu H C, Deb D, Liu H, Tang J L and Jain A K. 2020. Adversarial attacks and defenses in images, graphs and text: a review. International Journal of Automation and Computing, 17(2): 151-178 [DOI: 10.1007/s11633-019-1211-x]

Yuan T H, Ji S H, Zhang P C, Cai H B, Dai Q Y, Ye S J and Ren B. 2022. Adversarial example generation method for black box intelligent speech software. Journal of Software, 33(5): 1569-1586 (袁天昊, 吉顺慧, 张鹏程, 蔡涵博, 戴启印, 叶仕俊, 任彬. 2022. 针对黑盒智能语音软件的对抗样本生成方法. 软件学报, 33(5): 1569-1586) [DOI: 10.13328/j.cnki.jos.006549]

Zhang X, Zhao J B and LeCun Y. 2015. Character-level convolutional networks for text classification//Proceedings of 2015 Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems (NeurIPS 2015). Montreal, Canada: NIPS: 649-657

Zhang W E, Sheng Q Z, Alhazmi A and Li C L. 2020. Adversarial attacks on deep-learning models in natural language processing: a survey. ACM Transactions on Intelligent Systems and Technology, 11(3): #24 [DOI: 10.1145/3374217]

Zhao J J, Wang J W and Wu J F. 2023. Adversarial attack method identification model based on multi-factor compression error. Journal of Image and Graphics, 28(3): 850-863 (赵俊杰, 王金伟, 吴俊凤. 2023. 基于多质量因子压缩误差的对抗样本攻击方法识别. 中国图象图形学报, 28(3): 850-863) [DOI: 10.11834/jig.220516]

Zeng J H, Xu J H, Zheng X Q and Huang X J. 2023. Certified robustness to text adversarial attacks by randomized. Computational Linguistics, 49(2): 395-427 [DOI: 10.1162/coli_a_00476]

作者简介

陈宇涵,男,硕士研究生,主要研究方向为文本对抗样本、自然语言处理和对抗性检测。E-mail:chenyuhan@xmut.edu.cn

杜侠,通信作者,男,讲师,硕士生导师,主要研究方向为机器学习与计算机视觉、人工智能安全和语音识别。E-mail:duxia@xmut.edu.cn

王大寒,男,研究员,硕士生导师,主要研究方向为模式识别、计算机视觉、文本检测与识别。E-mail:wangdh@xmut.edu.cn

吴芸,女,教授,硕士生导师,主要研究方向为智能信息处理、数据挖掘、生物信息计算。E-mail:ywu@xmut.edu.cn

朱顺痣,男,教授,主要研究方向为数据挖掘、视频分析与处理、信息推荐和系统工程。E-mail:szzhu@xmut.edu.cn

严严,男,教授,主要研究方向为计算机视觉和模式识别。E-mail:yanyan@xmu.edu.cn