基于微调前后层相似度的对抗样本检测方法

陈佳冕[1],姚 迅[1],何 园[1],胡新荣[1],杨 捷[2]

(1. 武汉纺织大学 计算机与人工智能学院,湖北 武汉 430200;
2. 澳大利亚伍伦贡大学 工程与信息科学学院,新南威尔士州 伍伦贡 2522)

摘 要: 对抗样本通过微小扰动误导分类模型产生错误预测,发展对抗样本检测技术对提升模型鲁棒性至关重要。现有的对抗样本检测方法多利用模型末端输出层特征,而忽视了模型内部的层级化表征信息。分析了预训练语言模型中对抗样本与原始样本在各隐藏层输出的表征差异,并提出了分层对抗样本检测框架,通过计算模型微调前后原始样本与对抗样本每一层隐藏层输出的相似性,揭示两者在层间相似度变化模式上的显著差异。基于此,提取层相似度作为关键特征,并采用One-Class SVM模型学习正常样本的层相似度分布,进而实现对对抗样本的准确识别。将此检测框架在4个文本分类基准数据集和3种攻击方法下进行了实验,结果表明该方法在F1和AUC指标上均优于当前领域最优算法,较最强基线最高提升8.6%,且在多种预训练编码器架构上表现出良好的泛化能力。

关键词: 对抗样本检测;防御机制;隐藏层特征;模型鲁棒性;One-Class SVM

DOI: 10. 11907/rjdk. 251848 扫描二维码阅读全文:

中图分类号: TP391.41;TP18 **文献标识码:** A **文章编号:** 1672-7800(XXXX)0XX-0001-09

Adversarial Example Detection via Layer-wise Similarity before and after Fine-tuning

CHEN Jiamian[1], YAO Xun[1], HE Yuan[1], HU Xinrong[1], YANG Jie[2]

(1. *School of Computer Science and Artificial Intelligence*, *Wuhan Textile University*, *Wuhan* 430200, *China*;
2. *Faculty of Engineering and Information Sciences*, *University of Wollongong*, *Wollongong* 2522, *Australia*)

Abstract: Adversarial examples can mislead classification models into making incorrect predictions through minor perturbations, making the development of adversarial example detection technology crucial for enhancing model robustness. Existing detection methods predominantly rely on features from the model′s final output layer, leaving the hierarchical representational information within the model′s internal layers underutilized. This study analyzes the representational differences between adversarial examples and original samples in the hidden layer outputs of pre-trained language models and proposes a hierarchical adversarial example detection framework. By recording and computing the similarity of hidden layer outputs between original and adversarial samples before and after model fine-tuning, we find that the variation patterns of inter-layer similarity differ significantly between adversarial and original samples. Using layer-wise similarity as the core metric, a One-Class SVM model is trained to learn the distribution characteristics of layer similarity in normal samples, thereby effectively identifying adversarial examples. The proposed detection framework was evaluated on four benchmark text classification datasets under three attack methods. Experimental results show that it outperforms the current state-of-the-art algorithms in terms of both F1 and AUC scores, achieving up to an 8.6% improvement over the strongest baseline. Additionally, the framework demonstrates good generalization capabilities across different pre-trained encoder architectures.

Key Words: adversarial sample detection; defense mechanisms; intermediate layer features; model robustness enhancement; One-Class SVM

收稿日期: 2026-03-17

作者简介: 陈佳冕(2001-),女,武汉纺织大学计算机与人工智能学院硕士研究生,研究方向为自然语言处理;姚迅(1969-),男,武汉纺织大学计算机与人工智能学院讲师,研究方向为自然语言处理;何园(2001-),男,武汉纺织大学计算机与人工智能学院硕士研究生,研究方向为自然语言处理;胡新荣(1973-),女,武汉纺织大学计算机与人工智能学院教授,研究方向为图像处理与媒体计算;杨捷(1987-),男,澳大利亚伍伦贡大学工程与信息科学学院博士研究生,研究方向为自然语言处理。本文通讯作者:陈佳冕。

0 引言

自然语言处理(Natural Language Processing,NLP)领域因深度学习和预训练语言模型(Pre-trained Language Models, PLMs)如 BERT(Bidirectional Encoder Representations from Transformers)的兴起而取得巨大进步[1]。预训练语言模型在文本分类、情感分析等任务中表现出色,但它们通常对对抗攻击较为敏感。对抗攻击通过在正常样本中引入不易察觉的扰动,构造出对抗样本,导致模型产生错误输出[2]。针对该问题,现有研究主要遵循两条技术路线:①通过对抗性训练等方法提升模型鲁棒性;②在输入端采用对抗样本检测(Adversarial Example Detection,AED)识别并过滤对抗样本,以减轻其危害[3]。

现有的对抗样本检测方法主要分为3大类:基于扰动的方法、基于推理的方法和基于分布的方法。基于扰动的方法通过对输入施加如单词替换等细微扰动,测试模型的响应变化以识别对抗样本[4-6];基于推理的方法通过分析模型内部属性,如输入损失或输出不确定性,识别潜在对抗样本[7-9];基于分布的方法利用对抗样本常位于决策边界附近的特性进行检测[10]。

预训练语言模型的不同隐藏层能够捕捉不同抽象级别的语义特征,然而现有检测方法大多仅利用最终输出层的信息。鉴于此,本文提出了一种基于微调前后层相似度的对抗样本检测方法,通过多层表征之间的变化模式以有效识别对抗样本。

本文通过实验观察发现,对抗样本与原始样本在模型隐藏层输出上存在显著差异,说明对抗攻击不仅改变最终层的预测,也扰动了各隐藏层的表示。基于此,本文提出了一种基于隐藏层的检测框架,通过捕捉对抗攻击在隐藏层引入的异常扰动以实现检测。具体方法为:计算模型微调前后各隐藏层输出的相似度,验证其具有能反映原始样本和对抗样本层间特征变化规律的信息,并基于层相似度训练二分类器进行识别。

本文采用单类分类方法实现对抗样本检测,训练单类支持向量机 One-Class SVM[11]构建检测器,通过捕捉原始样本在多层表征上的相似度分布规律,从而有效区分不同攻击方式生成的对抗样本。由于不依赖于特定攻击类型信息,该框架具有良好的攻击泛化能力,同时在不同预训练编码器(如 BERT、RoBERTa)上也表现出稳定的检测性能,体现出该方法跨架构应用的鲁棒性与实用性。

本文提出的对抗样本检测框架在提升模型鲁棒性方面具有以下显著优势:①强泛化与高鲁棒性,不依赖攻击先验,可检测多类扰动,在4个数据集、3类攻击场景下,F1平均提升≥3.0%;②深层次的特征分析能力;③适应不同编码器架构,具备良好的跨架构迁移能力。

1 相关工作

在自然语言处理安全领域,对抗攻防是持续演进的博弈体系。对抗攻击通过字符、单词和句子级扰动生成对抗样本[12-14],防御则围绕扰动分析、推理统计与分布建模构建检测机制。近年来,随着对预训练模型内部表示机理的深入理解,隐藏层表征所蕴含的丰富语义与结构信息为对抗检测提供了新突破点。本文系统梳理相关研究脉络,并探讨基于隐藏层表示的防御新范式,为所提出的层次化检测框架奠定理论基础。

1.1 对抗攻击方法

对抗攻击通过在输入数据中添加微小扰动,使机器学习模型误判[15-16]。依据扰动粒度可分为3类:字符级、单词级[17]和句子级[2,18]。字符级攻击通过修改单个字符实现[15];单词级攻击[16,19]涉及对单词的插入、删除和位置改变;句子级攻击通过文本复述或插入句子实现,通常保持语义和语法正确。

1.2 对抗样本检测方法

对抗样本检测通过训练二分类器识别并过滤对抗样本,以实现模型防护。现有方法主要分为3类:基于扰动的方法[3-4]、基于推理的方法[7-8]和基于分布的方法[10]。

基于扰动的方法通过主动修改输入样本来探测对抗样本:通过替换、掩盖或增删字符/单词构造扰动变体,再比较原样本与扰动变体间的模型输出差异,若预测结果发生显著变化,则判定该样本可能为对抗样本。例如,UAPAD(Universal Adversarial Perturbations)会注入随机噪声,并在模型微调期间对其进行细化[3];FGWS(Frequency-Guided Word Substitutions)使用预先收集的原始样本和对抗样本之间的词频差异以替换单词[4];WDR(Word Deletion and Replacement)根据单词的意义对单词进行屏蔽[5];CHECKHARD 对输入文本进行词汇替换等转换以生成多个版本,利用模型预测结果差异检测对抗样本[6]。

基于推理的方法关注模型输出层面。ADDMU(Adversary Detection with Data and Model Uncertainty)利用数据和模型不确定性估计以检测远离决策边界的对抗样本[7]。例如,Sharpness从几何角度分析输入损失分布,发现对抗样本在其上呈现深且尖锐的局部最小值特性[8]。MLMD(Masked Language Model-based Detection)方法利用预训练的掩码语言模型,通过掩码和解掩码操作探测对抗样本与正常样本在数据流形上的变化,从而对二者加以区分[9]。

基于分布的方法依赖于特征分布,通过比较测试样本在特征空间上的概率密度以检测对抗样本。RDE(Robustness via Disturbance Detection)通过鲁棒密度估计以识别和过滤那些偏离正常数据分布的对抗样本[10]。

对抗攻击类型日益多样且数据分布动态变化,现有检测方法常面临泛化能力不足的挑战。为此,本文提出了基

于隐藏层相似度信息的检测路径,该方法不依赖于特定攻击先验或数据分布假设,通过挖掘模型内部层级表征的一致性规律,实现跨攻击类型与数据场景的稳定泛化,为文本对抗样本识别提供了新的解决方案。

1.3　隐藏层表示的潜力

预训练语言模型如BERT和GPT在Transformer架构上对海量数据集进行训练,推动了自然语言处理的革命性进步[20-21]。诸多研究致力于探究预训练语言模型各层功能和特点,挖掘隐藏层的丰富信息。

例如,Geva等[22]通过干预多头自注意力(Multi-Head Self-Attention Sublayer, MHSA)和MLP(Multilayer Perceptron)子层,发现自回归语言模型中事实关联存储于低层,经早期MLP子层富集后,由高层MHSA子层提取属性。Llama SLayer 8B强调浅层网络对知识注入的重要性,通过增强浅层并剪枝深层以优化模型[23]。LayerSkip利用早期层生成草稿词元,后期层验证修正,实现快速推理[24]。Zhang等[25]通过Shapley值量化各层贡献,结合层剪枝实验,揭示早期层对模型性能的关键作用。Sastry等[26]利用Gram矩阵表征隐藏层特征相关性,检测预测类别与活动模式的不一致性,识别分布外样本。

在防御对抗攻击领域,深度学习模型隐藏层相关研究日益重要。这些研究通过分析和调整隐藏层表示,显著提升了模型对恶意输入的鲁棒性。LaCL(Layer-agnostic Contrastive Learning)提出基于对比学习的框架,通过隐藏层学习特定表示并组合成单一表示,提高OOD检测的有效性[27]。BATER(Bayesian Adversarial Example Detector)利用贝叶斯神经网络模拟隐藏层输出分布,通过分布离散程度检测对抗样本[28]。LED(Layer-specific Editing)通过重新对齐关键安全层抵御越狱攻击[29]。IMO(Invariant Features Masks for Out-of Distribution Text Classification)通过学习稀疏掩码层去除无关特征,提高OOD文本分类的泛化能力[30]。BLOOD(Between Layer Out-Of-Distribution)基于隐藏层表示转换的平滑度,利用分布内数据的平滑特性检测OOD数据[31]。

2　基于微调前后层相似度检测

本文提出了基于隐藏层的对抗样本检测框架(Detection via Layer-wise Similarity before and after Fine-tuning, DLSF),在阐述层相似度计算方法的基础上,可视化并分析了模型微调前后原始样本和对抗样本层相似度差异。并且,说明了如何将层相似度特征输入One-class SVM模型以实现对抗样本检测。

2.1　层相似度计算方法

用原始的数据集微调BERT分类模型,并保留微调前后的模型参数。设$x \in X$为输入样本,BERT模型共有L层。使用微调后的BERT模型$BERT_f$对每一条原始样本和对抗样本进行预测,则第l层的隐藏层输出可以表示$H_{f,l}(x)$。同样,使用微调前的BERT模型$BERT_p$对x进行预测,得到第l层的隐藏层输出$H_{p,l}(x)$。

对每个隐藏层输出向量进行L2归一化处理,并计算每条数据对应的微调前后的两个隐藏层输出之间的余弦相似度$s_l, l \in \{1, 2, \cdots, L\}$。数据$x$在模型微调前后第$l$层输出的相似度值如式(1)所示。

$$s_l = \frac{H_{f,l}(x) \cdot H_{p,l}(x)}{\left\| H_{f,l}(x) \right\| \cdot \left\| H_{p,l}(x) \right\|} \tag{1}$$

其中,$\cdot$表示向量的点积,$\|\cdot\|$表示向量的L2范数。

对于每条输入数据x,本文通过层相似度计算方法得到由每一层的相似度s_l,各层数据组成相似度向量$F(x)$,如式(2)所示。

$$F(x) = [s_1, s_2, \cdots, s_L] \tag{2}$$

$F(x)$即为所获取每个输入x的模型微调前后的层相似度特征向量,其中L是模型总层数。

2.2　模型微调前后原始样本与对抗样本层相似度差异

本文比较了BERT模型微调前的模型$BERT_p$和微调后的模型$BERT_f$对原始样本和对抗样本的层相似度,旨在捕捉模型在微调过程中对输入样本的敏感性,并量化原始样本和对抗样本的差异。

首先对BERT模型的全部12层进行了微调,微调前后模型层数相同,但参数不同:微调前的模型参数是预训练阶段学习到的通用语言模型参数,捕捉语言通用特征;微调后的模型参数针对特定文本分类任务进行了调整,以更好地适应任务数据分布。

图1左边的虚线框是层相似度计算模型,图中输入x表示输入样本,将其送入微调前的BERT模型$BERT_p$和微调后的模型$BERT_f$,分别得到它们每一层隐藏层的输出$H_{p,l}$和$H_{f,l}$。其中,$l \in \{1, 2, \cdots, L\}$,表示各隐藏层序号,$L$表示模型总层数,$s_l$表示通过计算$H_{p,l}$和$H_{f,l}$的余弦相似度得到的第$l$层的层相似度,将所有层的层相似度$[s_1, s_2, \cdots, s_L]$输入检测器$g(\cdot)$,输出为对抗样本或原始样本。

在SST-2[32]和AGNews[33]数据集上,TextFooler[16]、TextBugger[19]、DeepWordBug[15]3种攻击方法下原始样本和对抗样本隐藏层各层相似度实验结果如图2所示。图中显示了第1层到第L层相似度值变化曲线,其中▮表示原始样本的层相似度,▲表示对抗样本的层相似度,横轴表示层数,纵轴表示相似度值,实线表示均值,阴影区域表示标准差。

图2显示,随着隐藏层层数增加,微调前后输出的层相似度总体呈下降趋势。浅层(1、2、3、4层)中,原始样本与对抗样本的相似度均较高且差异较小;自中间层(5、6、7、8、9层)起,二者差异开始扩大。这是因为模型在中间层逐渐编码上下文与语义信息,微调使其向任务目标空间对齐,原始样本包含更多任务相关信息,隐藏层表示变化较大,相似度下降较快;而对抗样本受扰动影响,语义信息被

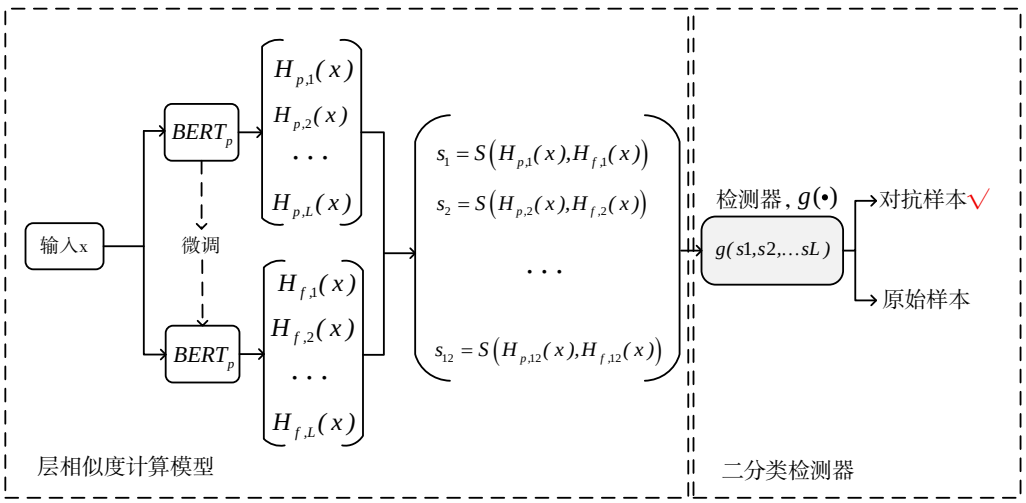

Fig. 1 Detection model based on layer-wise similarity

图 1 基于层相似度的检测模型

削弱，微调对其表示影响较小，因此相似度下降平缓。在深层（10、11、12层），原始样本相似度急剧下降，对抗样本则下降较缓，这是由于深层提取的是与任务高度相关的判别性特征，微调对其调整显著；而对抗扰动干扰了模型对任务信息的学习，使微调前后对抗样本的深层表示变化较小，相似度曲线较为平缓。

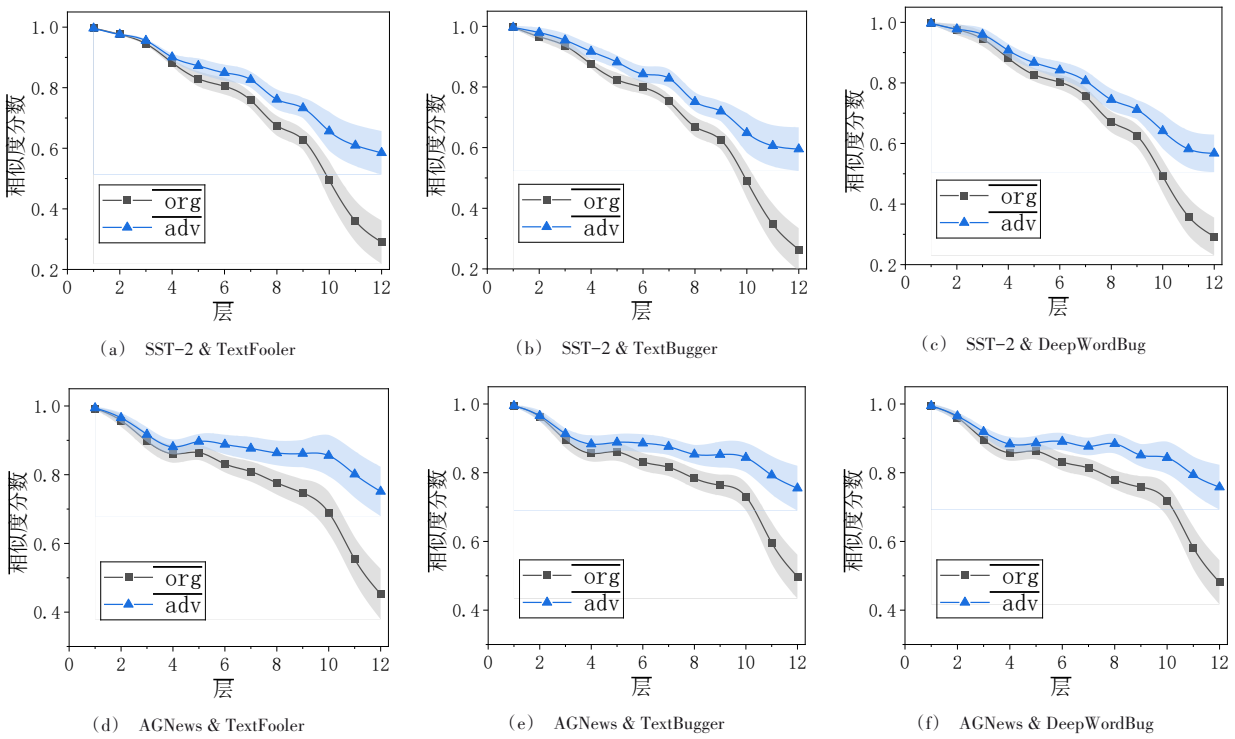

Fig. 2 Similarity evolution across hidden layers before and after fine-tuning on SST-2 and AGNews under character-level and word-level attacks

图 2 SST-2和AGNews数据集在字符级和单词级攻击下微调前后隐藏层相似度变化

这种结果符合本文对对抗样本在模型训练中的表现预期，也为利用该特性对对抗样本进行检测提供了直观依据。这种现象不仅存在于AGNews和SST-2数据集上的TextAttack攻击，还存在于其它数据集和攻击方法中[17]。对此，本文实验部分将给出更多验证结果。

基于上述观察，本文提出了一种新的检测算法：先计算模型微调前后所有隐藏层输出的相似度，如图1左边虚线框内容所示，再将层相似度向量$F(x)$作为特征输入为后续One-Class SVM用来检测识别对抗样本，如图1两个虚线框所示。

2.3 One-Class SVM检测器框架

在上文研究中，本文提取了层相似度特征$F(x)$。接下来，选用单类支持向量机One-Class SVM作为检测器，检测器的目标是根据输入$F(x) \in F$判断输入样本是否属于原始且未被攻击的样本分布[11]。

One-Class SVM定义了一个特征映射ϕ：

$F \subset \mathbb{R}^d \to S \subset \mathbb{R}^h$。若训练集中有 m 个样本,将输入样本即 d 维特征向量 $\{F(x_i)\}_{i=1}^m \in \mathbb{R}^d$ 映射到高维 h 维特征空间 S。目标是在特征空间 S 中找到一个最优的分离超平面,将所有输入样本与原点分离,同时最大化到原点的距离。

本文选用高斯径向基函数(Gaussian Radial Basis Function, RBF)作为核函数,如式(3)所示。

$$k\left(F\left(x_i\right), F\left(x_j\right)\right) = \exp\left(-\gamma \left\| F\left(x_i\right) - F\left(x_j\right)\right\|^2\right) \tag{3}$$

其中,γ 为核宽度参数,用于控制核函数的宽度,影响模型对数据的拟合程度。

One-Class SVM 的训练过程涉及一个对偶二次规划(Quadratic Programming, QP)问题。利用拉格朗日乘数,优化问题可表示为如式(4)所示。

$$\begin{gathered} \min_{\alpha} \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j k\left(F\left(x_i\right), F\left(x_j\right)\right) \\ s.t. 0 \leqslant \alpha_i \leqslant \frac{1}{vm}, \\ \sum_i \alpha_i = 1 \end{gathered} \tag{4}$$

其中,α_i 是拉格朗日乘数,$\nu \in (0, 1]$ 是训练误差的上界。

给定一个输入样本 $F(x) = [s_1, s_2, \cdots, s_L]$,其与分类超平面的有符号距离通过式(5)的决策函数加以计算。

$$d\left(F(x)\right) = \sum_i \alpha_i k\left(F\left(x_i\right), F(x)\right) - \rho \tag{5}$$

其中,α_i 是拉格朗日乘数,偏移量 ρ 可通过 $\rho = \sum_j \alpha_j k(F\left(x_j\right), F(x_i))$ 恢复。$d(F(x))$ 的值越大表示样本越接近正常样本分布,负值则指示异常。因此,本文设置决策规则为:若 $d(F(x)) < \delta$ 则判定 x 为对抗样本;否则为原始样本。阈值 δ 通过在验证集上优化 F1 分数而确定。

3　实验与结果分析

3.1　实验设置

3.1.1　数据集

实验在 4 个文本分类数据集上进行,这 4 个数据集的训练样本、测试样本数量和数据平均长度信息如表 1 所示。①Stanford Sentiment Treebank 2(SST-2)[32]:使用斯坦福大学情感树库语料库进行情感分析的公认资源;②Yelp Reviews(YELP)[33]:Yelp 平台的评论数据集,包含业务评级和客户反馈;③AG's News Corpus(AGNews)[33]:包含世界、体育、商业和科技 4 个主题的新闻文章数据集;④Internet Movie Database(IMDB)[34]:从 IMDB 平台收集的情感分类数据集。

3.1.2　攻击算法

实验种使用 TextAttack[17] 实现了 3 种对抗性攻击方法以扰动输入,包括:①TextBugger[19] 执行空格插入、字符删除/交换和同义词替换的扰动;②DeepWordBug[15] 删除、替换和插入字符到输入;③TextFooler[16] 用同义词替换重要单词。其中,TextBugger 和 DeepWordBug 方法主要集中在字符级攻击,而 TextFooler 方法主要用于单词级攻击。实验统一约束扰动粒度、文本语义相似度等超参,将 3 种攻击在全部数据集上的攻击成功率(Attack Success Rate, ASR)差值控制在 5% 以内,消除攻击难度不一致带来的对比偏差。

Table 1　Benchmark statistics

表 1　基准统计数据

数据集	训练集/条	测试集/条	平均长度/字符数
SST-2	67 300	1 821	16
YELP	560 000	38 000	152
AGNews	120 000	7 600	50
IMDB	25 000	25 000	161

3.1.3　实验细节

对于所提出的相似度检测 DLSF 方法,本文实验在 GPU 为 NVIDIA RTX 4090、Python 3.7.0、PyTorch 1.7.0+cu110 的环境下进行,采用 BERT-base 作为上下文编码器[1]。训练时对模型全部 12 层执行反向传播,学习率与批量大小参考最近微调工作的常见设置[35],优化器为 Adam(学习率 2e-5),批量大小为 32;最大序列长度 512 能覆盖本文所用数据集的文本长度,因此输入序列最大长度设置为 512。模型在 10 个 epoch 内即可收敛,因此将训练上限设为 10 个 epoch。

为了进行公平比较,本文采用了与文献[7]相同的推理过程。从原始的测试数据集和被攻击之后的对抗测试数据集中分别随机选择 t 个样本。对于所使用的 SST-2、YELP、AGNews 和 IMDB 数据集,t 分别设置为 2 000、4 000、4 000 和 4 000。相应地,采用检测方法以区分正常和被攻击的测试样本。

3.1.4　评估指标

参照已有研究,本文实验采用了两种测量方法以评估对抗性检测的有效性[5,7]。具体而言,F1 评估了模型在精确度和召回能力之间的平衡;曲线下面积(AUC)提供了模型在所有可能分类阈值上的性能综合度量。因此,F1 和 AUC 的值越高,表明检测准确度越好。

3.2　实验结果

本文实验首先在干净的训练数据集上微调一个 BERT 分类模型,然后通过 TextAttack 扰动微调后的模型以产生对抗样本。实验所用对抗样本均为攻击成功的扰动样本,所有对比检测方法均基于相同攻击强度的对抗样本集开展评估,从样本层面规避了对比偏差,确保实验结果的公平性与可比性。

所提出的方法与几种目前最先进的方法进行了比较,包括 SCAD[35]、TextDefense[36]、UAP[3]、Sharpness[8]。其中,SCAD 基于子空间聚类检测偏离低维子空间的样本;TextDefense 基于词重要性分数分散度检测;UAP 通过注入通用扰动并观察响应变化;Sharpness 利用对抗样本在损失 Landscape 的尖锐特性。上述方法均需修改输入或计算梯

度，而本文仅利用隐藏层表示差异，在检测阶段无需反向传播或扰动输入，也无需知道攻击类型。

基于层相似度的检测方法DLSF在字符级和单词级攻击下，使用随机种子进行了5次运行，其平均结果如表2所示。

表2 Performance comparison among DLSF and current SOTAs

表2 DLSF与当前最优方法(SOTAs)性能比较 (%)

数据集	检测方法	TextFooler		TextBugger		DeepWordBug	
		F1	AUC	F1	AUC	F1	AUC
SST-2	UAP	82.1	94.0	85.2	93.9	77.8	91.5
	Sharpness	86.3	92.8	89.9	93.9	80.7	90.7
	SCAD	74.6	86.8	80.2	90.7	73.3	87.9
	TextDefense	75.4	87.0	74.8	85.5	71.0	84.1
	DLSF	92.0±0.4	94.2±0.7	92.3±0.5	95.3±1.1	89.3±0.4	91.6±1.2
IMDB	UAP	74.9	77.8	72.5	74.6	73.5	75.7
	Sharpness	86.8	95.3	86.3	91.0	86.2	91.7
	SCAD	94.4	98.2	93.4	97.7	93.8	97.1
	TextDefense	88.2	91.2	88.7	92.5	87.4	90.6
	DLSF	97.1±0.7	98.6±0.2	97.1±0.7	97.9±0.5	96.9±0.6	97.5±0.9
YELP	UAP	87.4	90.1	85.7	90.5	86.7	90.7
	Sharpness	92.9	94.2	91.1	94.2	90.5	94.8
	SCAD	93.5	96.3	91.5	95.4	93.1	96.1
	TextDefense	91.6	95.3	91.2	94.8	90.1	93.1
	DLSF	97.6±0.6	97.7±0.6	92.6±0.2	97.6±0.7	97.6±1.0	97.1±0.5
AGNews	UAP	94.6	95.2	92.5	94.7	93.8	94.8
	Sharpness	94.7	95.8	92.9	95.7	93.8	95.7
	SCAD	95.6	97.0	93.7	96.3	94.3	96.1
	TextDefense	92.9	94.7	92.6	95.7	92.3	94.1
	DLSF	96.3±0.8	97.2±0.7	95.7±0.5	96.5±0.4	96.8±0.3	97.2±0.5

从上述实验结果看，基于层相似度的检测方法与所有基线方法相比都表现更佳，在4个数据集、3种攻击方法下展现了良好性能。例如，在SST-2数据集上，与基线方法(Sharpness)相比，本文方法在DeepWordBug攻击下F1得分提升了8.6%，在TextFooler攻击下F1得分提升了5.7%；在IMDB数据集上，与基线方法(SCAD)相比，TextBugger攻击下F1得分超过最强基线3.7%，在YELP数据集上，DeepWordBug攻击下超过最强基线(SCAD)4.5%。这些结果充分验证了本文方法的优秀表现，也表明了该方法的有效性和可行性。

为了全面评估DLSF的综合性能，本文系统对比了所提方法与主流基线在训练推理速度上的表现，使用了GPU为NVIDIA RTX 4090的机器进行效率评测。表3显示了在TextFooler攻击下使用SST-2和IMDB数据集在数据集训练1 000个样本并检测1 500个样本所需的GPU时间。

效率对比结果表明，DLSF在训练阶段和测试阶段均显著优于所有对比方法。该优势主要得益于其轻量化的特征提取设计：仅需单次前向传播即可获取多层级相似度特征，避免了复杂的迭代计算与后处理流程。DLSF在精度与效率的平衡上取得了突破，为对抗样本的实时在线检测提供了更具实用价值的解决方案。

Table 3 Computational efficiency calculate of each detection method in seconds

表3 以秒为单位计算各检测方法计算效率 (s)

检测方法	训练时间		测试时间	
	SST-2	IMDB	SST-2	IMDB
TextDefense	58.0	69.0	18.7	21.5
SCAD	34.4	45.7	12.1	12.6
DLSF	27.9	35.1	9.7	10.1

3.3 消融研究

3.3.1 基于RoBERTa编码器的对抗样本检测

为验证层相似度检测方法(DLSF)对不同基础编码器的兼容性，本文进一步评估了其在不同编码器结构下的性能表现。具体而言，使用RoBERTa编码器替换原有的BERT编码器[37]。并且，在YELP与AGNews数据集上使用TextAttack公开攻击模型生成对抗样本，以确保实验环境一致[17]。选取现有方法中表现最优的SCAD与TextDefense作为基线方法进行对比。

本文方法和基线方法之间的性能比较如表4所示。结果显示，DLSF在使用RoBERTa编码器时，在所有数据集的F1分数上均显著优于SCAD和TextDefense。这进一步证实了DLSF方法在不同底层编码器(RoBERTa或BERT)上的稳定性和有效性，突出了其鲁棒性和泛化能力。

Table 4 Impact analysis of the encoder (RoBERTa) on adversarial detection

表4 底层编码器RoBERTa对对抗性检测的影响分析 (%)

数据集	检测方法	TextFooler		TextBugger		DeepWordBug	
		F1	AUC	F1	AUC	F1	AUC
YELP	SCAD	94.3	97.1	92.0	97.6	93.1	97.1
	TextDefense	92.1	96.1	91.7	95.6	91.1	95.1
	DLSF	96.8±0.7	97.2±0.3	93.1±1.2	97.9±0.7	94.2±0.4	97.3±0.7
AGNews	SCAD	95.9	97.1	93.4	96.5	94.5	96.2
	TextDefense	93.9	95.1	91.4	94.5	92.5	94.2
	DLSF	96.4±0.9	98.2±0.6	95.8±0.4	97.5±0.6	94.6±0.7	97.3±0.6

3.3.2 基于有监督分类器的对抗样本检测

为了验证所提出方法在有监督检测任务中的性能，本文选取层相似度作为判别特征，构建了基于多层感知机(MLP)的有监督二分类器，并命名为有监督层相似度检测模型S_DLSF。该模型旨在通过学习层相似度特征，实现对对抗样本与原始样本的区分。表5展示了其在SST-2和IMDB数据集上的检测成果，涵盖F1值与AUC值，并纳入当前最优的有监督对抗样本检测方法WDR[5]和MLMD[9]以作对比。

实验数据显示，S_DLSF在SST-2与IMDB数据集和3种攻击方法上均稳定优于WDR和MLMD，尤其在SST-2数据集DeepWordBug攻击下F1达到89.2%，AUC达到96.2%，领先基线方法超过5.0%。这凸显了正常样本与对抗样本在层相似度所含信息下具有显著可区分性，该特性使不同架构模型或分类器能够有效分离二者，意味着此方法既适

用于无监督检测，也适配于有监督检测场景。

Table 5　Detection results of the supervised MLP-based layer-similarity method (S-DLSF)

表5　基于有监督的MLP分类器的层相似度方法(S_DLSF)检测结果

(%)

Dataset	Methods	TextFooler		TextBugger		DeepWordBug	
		F1	AUC	F1	AUC	F1	AUC
SST-2	WDR	73.4	82.6	77.8	88.4	63.6	84.1
	MLMD	90.3	92.5	85.5	94.1	83.9	90.7
	S_DLSF	92.3±0.4	96.5±0.2	91.1±0.5	95.7±0.4	89.2±0.6	96.2±0.7
IMDB	WDR	90.2	91.2	87.3	89.7	87.6	90.1
	MLMD	94.7	96.2	95.6	96.3	93.6	96.2
	S_DLSF	94.8±0.5	97.8±0.6	96.8±0.3	98.2±0.2	96.6±0.5	98.5±0.4

3.3.3　基于单层层相似度的对抗样本检测

为了评估每一层隐藏层相似度在对抗样本检测中的有效性，本文进行了基于单层的消融实验，分别单独使用BERT模型的第1层至第12层的隐藏层相似度作为特征，在IMDB与AGNews数据集上进行检测，计算相应的F1值和AUC值。图3给出了两种数据集在不同攻击方法下各单层相似度检测的F1和AUC数据，同时给出本文方法(记为m)的性能以供对比。可以看出，随着层数增加，IMDB和AGNews单层检测的F1值和AUC值总体呈上升趋势，尤其在第10层至第12层提升显著，表明深层特征更有利于捕捉语义信息以识别对抗扰动。例如，IMDB数据集上单层消融实验显示第12层在TextBugger攻击下取得最高F1=82.97%和AUC=91.14%，但其性能仍显著低于使用全部12层相似度的完整模型F1=97.1%和AUC=97.9%。结合图2实验结果证明，仅用单层只能捕获某一深度的特征，而全12层相似度向量完整刻画了扰动在模型内部的传播轨迹，这种使用所有层相似度的模式比仅使用单层检测更具判别力和稳健性。

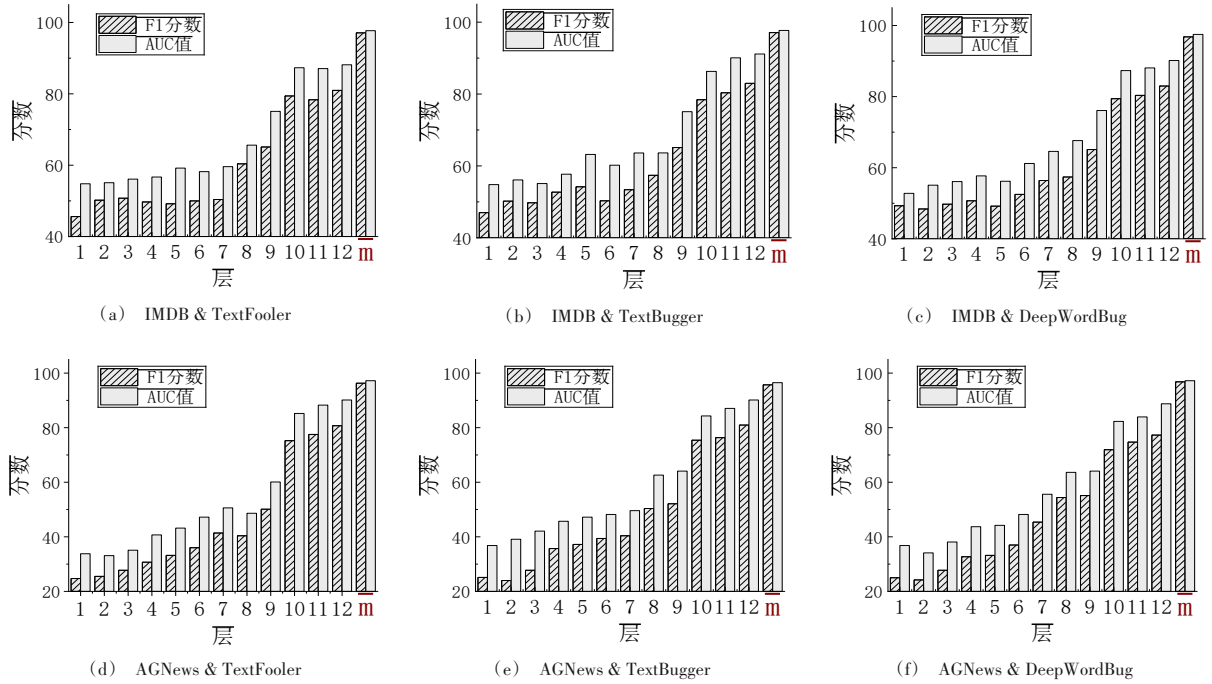

Fig. 3　Single-layer detection using layer-similarity on the IMDB and AGNews datasets

图3　基于IMDB和AGNews数据集的层相似度单层检测

3.3.4　基于一阶差分的对抗样本检测

为验证直接拼接层相似度这一特征构建方式的有效性，同时探索差分变换等策略对检测性能的提升价值，本文以层相似度序列为基础，设计一阶差分变换这种特征构建方式，从定量角度对比分析基于一阶差分的层相似度检测方法DLSF-Diff和原始直接拼接方式DLSF的效果。其中，各层相似度记为$\{s_l\}_{l=1}^{L}$，其中L为模型隐藏层总数。

一阶差分通过计算相邻层相似度的差值以量化层间相似度的变化速率，定义如式(6)所示。

$$\Delta s_l = s_{l+1} - s_l \tag{6}$$

其中，$l = 1, 2, \cdots, L-1$，Δs_l表示第l层到第$l+1$层的相似度变化量。差分特征向量如式(7)所示。

$$F_{diff}(x) = [\Delta s_1, \Delta s_2, \cdots, \Delta s_{L-1}] \tag{7}$$

而原方法DLSF采用的直接拼接特征向量为$F(x) = [s_1, s_2, \cdots, s_L]$，二者作为核心对比特征开展消融实验。

选取SST-2、IMDB两个数据集，在TextFooler、DeepWordBug两种典型攻击下开展实验。实验设置与3.1节保持一致，以F1和AUC为评估指标，结果如表6所示。

结果显示，DLSF在所有数据集与攻击方法下的F1、AUC指标均优于DLSF-Diff，充分验证了原方法直接拼接层相似度特征设计的最优性与有效性。DLSF方式完整保留了各层相似度的绝对数值与序列演化信息，能够精准捕捉原始样本与对抗样本的核心分布差异，是兼顾检测性能与计算效率的合理选择。

Table 6 Ablation comparison of detection performance of different feature construction methods

表6 不同特征构建方式的检测性能消融比较 (%)

Dataset	Methods	TextFooler		DeepWordBug	
		F1	AUC	F1	AUC
SST-2	DLSF-Diff	90.8±0.5	93.5±0.8	88.5±0.6	90.9±0.9
	DLSF	92.0±0.4	94.2±0.7	89.3±0.4	91.6±1.2
IMDB	DLSF-Diff	96.2±0.8	97.9±0.3	96.0±0.7	96.8±0.8
	DLSF	97.1±0.7	98.6±0.2	96.9±0.6	97.5±0.9

3.3.5 基于组合层的层相似度对抗样本检测

为探究模型不同深度层级在对抗样本检测中的相对重要性，并验证全层特征融合策略的必要性，本文设计了层级组合消融实验。通过系统地分组评估浅层（1、2、3、4层）、中层（5、6、7、8、9层）、深层（10、11、12层）以及跨层组合的检测性能。在YELP和IMDB数据集上开展实验，使用F1和AUC作为评估指标，结果如表7所示。

Table 7 Combined layer detection based on layer similarity

表7 基于层相似度的组合层检测 (%)

层	Textfooler				Textbugger			
	YELP		IMDB		YELP		IMDB	
	F1	AUC	F1	AUC	F1	AUC	F1	AUC
浅层	30.4	56.5	48.4	57.7	39.6	59.8	50.7	58.6
中层	54.6	61.2	67.0	75.5	54.7	65.9	66.3	74.1
深层	82.6	91.3	81.9	89.3	86.8	90.4	84.2	90.1
浅层+深层	89.5	94.5	89.5	95.5	92.3	97.0	93.8	95.3
浅层+中层	58.5	74.1	69.1	79.5	53.5	65.5	67.4	76.4
中层+深层	90.0	94.8	91.6	95.2	92.4	97.6	94.0	95.5
所有层	97.6	97.7	92.6	97.6	97.1	98.6	97.1	97.9

结果表明，深层特征是检测性能的核心，其F1分数超过80%，显著优于中层和浅层，证实对抗扰动在高层语义中引发关键判别性偏移。“浅层+深层”与“中层+深层”组合性能接近“所有层”，说明深层特征与中/浅层特征具强互补性。而“浅层+中层”组合性能一般，证明缺乏深层语义抽象则难以有效检测。

综上，“所有层”组合凭借完整层级演化轨迹实现了最优且最稳定的性能，验证了全层特征融合对捕捉对抗扰动完整传播模式的完备性与鲁棒性价值。

4 结语

本文提出了一种基于层相似度的对抗样本检测框架，通过对比模型在微调前后对原始样本与对抗样本的隐藏层输出相似性差异，实现对对抗样本的识别。该方法采用One-Class SVM作为分类器，具有无需依赖特定攻击假设、适用性广的特点。实验表明，所提方法在鲁棒性与泛化能力方面显著优于当前主流检测算法，能够有效应对多种常见对抗攻击，且具备较高的检测精度与较低的计算开销。为提升检测效率与特征判别力，后续研究中可探索引入注意力机制动态选择关键隐藏层进行相似度计算，并计划将其扩展至更多元、复杂的攻击环境中加以验证与优化。

参考文献：

[1] DEVLIN J, CHANG M W, LEE K. BERT: pre-training of deep bidirectional Transformers for language understanding [C]//Minneapolis: The 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019.

[2] WANG Y, WANG C, ZHAN J. Text FCG: fusing contextual information via graph learning for text classification [J]. Amsterdam: Expert Systems with Applications, 2023(c): 119658.

[3] GAO S, DOU S, ZHANG Q. On the universal adversarial perturbations for efficient data-free adversarial detection [C]//Toronto: 61st Annual Meeting of the Association for Computational Linguistics, 2023.

[4] MOZES M, STENETORP P, KLEINBERG B. Frequency-guided word substitutions for detecting textual adversarial examples [C]//Online: The 16th Conference of the European Chapter of the Association for Computational Linguistics, 2021.

[5] MOSCA E, AGARWAL S, RANDO J. That is a suspicious reaction!: interpreting logits variation to detect NLP adversarial attacks [C]//Dublin: The 60th Annual Meeting of the Association for Computational Linguistics, 2022.

[6] NGUYEN-SON H Q, UNG H Q, HIDANO S. CheckHARD: checking hard labels for adversarial text detection, prediction correction, and perturbed word suggestion [C]//Abu Dhabi: The 2022 Conference on Empirical Methods in Natural Language Processing, 2022.

[7] YIN F, LI Y, HSIEH C J. ADDMU: detection of far-boundary adversarial examples with data and model uncertainty estimation [C]//Abu Dhabi: The 2022 Conference on Empirical Methods in Natural Language Processing, 2022.

[8] ZHENG R, DOU S, ZHOU Y. Detecting adversarial samples through sharpness of loss landscape [C]//Toronto: 61st Annual Meeting of the Association for Computational Linguistics, 2023.

[9] ZHANG X, ZHANG Z, ZHONG Q. Masked language model based textual adversarial example detection [C]//New York: The 2023 ACM Asia Conference on Computer and Communications Security, 2023.

[10] YOO K Y, KIM J, JANG J. Detection of adversarial examples in text classification: Benchmark and baseline via robust density estimation [C]// Dublin: 60th Annual Meeting of the Association for Computational Linguistics, 2022.

[11] OLKOPF B S, WILLIAMSON R, SMOLA A. Support vector method for novelty detection [C]//Cambridge: Conference on Neural Information Processing Systems (NeurIPS), 1999.

[12] FANG Y S, CHEN Z H, HE K. Reinforced textual adversarial attack via automatic dilution [J]. Journal of Nanjing University (Natural Science), 2024, 60(6): 900-907.
房钰深，陈振华，何琨. 基于自动稀释的文本对抗攻击强化方法[J]. 南京大学学报(自然科学)，2024，60(6)：900-907.

[13] XIONG X, LIU Z R, ZHANG S. TextSwindler: a hard-label black-box textual adversarial attack algorithm [J]. Journal of Chinese Information Processing, 2024, 38(12): 18-29.
熊熙，刘钊荣，张帅. TextSwindler：面向硬标签黑盒文本的对抗攻击算法[J]. 中文信息学报，2024，38(12)：18-29.

[14] LUO S L, CHENG Y, WAN Y W. Black-box adversarial example attack against text-classification models with few queries [J]. Journal of Beijing Institute of Technology, 2024, 44(12): 1277-1286.
罗森林，程瑶，万韵伟. 少查询文本分类模型的黑盒对抗样本攻击[J]. 北京理工大学学报，2024，44(12)：1277-1286.

[15] GAO J, LANCHANTIN J, SOFFA M L. Black-box generation of adversar-

ial text sequences to evade deep learning classifiers[C]//San Francisco: 2018 IEEE Security and Privacy Workshops(SPW),2018.

[16] JIN D,JIN Z,ZHOU J T,et al. Is bert really robust? a strong baseline for natural language attack on text classification and entailment[C]//Hong Kong:The AAAI Conference on Artificial Intelligence,2020.

[17] MORRIS J X,LIFLAND E,YOO J Y. TextAttack: a framework for adversarial attacks, data augmentation, and adversarial training in NLP[C]//Online:The 2020 Conference on Empirical Methods in Natural Language Processing:System Demonstrations,2020.

[18] CHEN Y,GAO H,CUI G. From adversarial arms race to model-centric evaluation: Motivating a unified automatic robustness evaluation framework[C]//Toronto:61st Annual Meeting of the Association for Computational Linguistics,2023.

[19] LI J,JI S,DU T. TextBugger: generating adversarial text against real-world applications [C]//San Diego:Network and Distributed Systems Security Symposium,2019.

[20] RADFORD A,NARASIMHAN K,SALIMANS T. Improving language understanding by generative pre-training[C]//Vancouver:International Conference on Learning Representations,2018.

[21] VASWANI A,SHAZEER N,PARMAR N. Attention is all you need[C]//Long Beach: Conference on Neural Information Processing Systems(NeurIPS),2017.

[22] GEVA M,BASTINGS J,FILIPPOVA K. Dissecting recall of factual associations in auto-regressive language models [C]//Singapore: The 2023 Conference on Empirical Methods in Natural Language Processing,2023.

[23] CHEN T,TAN Z,GONG T. Llama SLayer 8B:shallow layers hold the key to knowledge injection[C]//Florida:Findings of the Association for Computational Linguistics,2024.

[24] ELHOUSHI M,SHRIVASTAVA A,LISKOVICH D. LayerSkip: enabling early exit inference and self-speculative decoding [C]//Bangkok: The 62nd Annual Meeting of the Association for Computational Linguistics, 2024.

[25] ZHANG Y,DONG Y,KAWAGUCHI K. Investigating layer importance in large language models [C]//Florida: The 7th Blackbox NLP Workshop: Analyzing and Interpreting Neural Networks for NLP,2024.

[26] SASTRY C S, OORE S. Detecting out-of-distribution examples with gram matrices [C]//Onilne: International Conference on Machine Learning,2020.

[27] CHO H,PARK C,KANG J. Enhancing out-of-distribution detection in natural language understanding via implicit layer ensemble[C]//Abu Dhabi:The 2022 Conference on Empirical Methods in Natural Language Processing,2022.

[28] LI Y,TANG T,HSIEH C J. Adversarial examples detection with Bayesian neural network[J]. IEEE Transactions on Emerging Topics in Computational Intelligence,2024,8(5):3654-3664.

[29] ZHAO W,LI Z,LI Y G. Defending large language models against jailbreak attacks via layer-specific editing[C]//Florida: Conference on Empirical Methods in Natural Language Processing,2024.

[30] FENG T,QU L Z,LI Z. IMO: greedy layer-wise sparse representation learning for out-of-distribution text classification with pre-trained models [C]//London:The 62nd Annual Meeting of the Association for Computational Linguistics,2024.

[31] JELENIĆ F,JUKIĆ J,TUTEK M. Out-of-distribution detection by leveraging between-layer transformation smoothness[C]//Macau:The International Joint Conference on Artificial Intelligence,2024.

[32] SOCHER R,PERELYGIN A,WU J Y. Recursive deep models for semantic compositionality over a sentiment treebank [C]//Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, 2013:1631-1642.

[33] ZHANG X,ZHAO J,LECUN Y. Character-level convolutional networks for text classification[DB/OL]. https://arxiv. org/abs/1509. 01626.

[34] MAAS A L,DALY R E,PHAM P T. Learning word vectors for sentiment analysis[C]//Portland: Annual Meeting of the Association for Computational Linguistics,2011.

[35] HU X,CHEN W,YANG J,et al. SCAD:subspace clustering based adversarial detector[C]//New York:The 17th ACM International Conference on Web Search and Data Mining,2024.

[36] SHEN L,PU Y,ZHANG X,et al. TextDefense: adversarial text detection based on word importance score dispersion[J]. IEEE Transactions on Dependable and Secure Computing,2025,22(4):4671-4684.

[37] LIU Y,OTT M,GOYAL N,et al. RoBERTa: a robustly optimized BERT pretraining approach[C]//Florence:The 57th Annual Meeting of the Association for Computational Linguistics,2019.

(责任编辑:孙　娟)