面向自然语言处理领域的对抗样本生成方法

张 影,方贤进,杨高明
(安徽理工大学 计算机科学与工程学院,安徽 淮南 232001)

摘 要:利用深度神经网络实现自然语言处理领域的文本分类任务时,容易遭受对抗样本攻击,研究对抗样本的生成方法有助于提升深度神经网络的鲁棒性。因此,提出了一种单词级的文本对抗样本生成方法。首先,设计单词的重要性计算函数;然后,利用分类概率查找到单词的最佳同义替换词,并将两者结合确定单词的替换顺序;最后,根据替换顺序生成与原始样本接近的对抗样本。在自然语言处理任务上针对卷积神经网络、长短时记忆网络和双向长短时记忆网络模型进行的实验表明:生成的对抗样本降低了模型的分类准确率和扰动率,且经过对抗训练之后模型的鲁棒性有所提高。

关键词:自然语言处理;文本对抗样本;文本分类;深度学习;单词级
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1673-629X(2023)03-0098-07
doi:10.3969/j.issn.1673-629X.2023.03.015

Adversarial Examples Generation Method for Natural Language Processing

ZHANG Ying,FANG Xian-jin,YANG Gao-ming
(School of Computer Science and Engineering,Anhui University of Science and Technology,Huainan 232001,China)

Abstract:When using deep neural networks to implement text classification tasks in the field of natural language processing,it is easy to be attacked by adversarial samples. Therefore,studying the generation method of adversarial samples can help improve the robustness of deep neural networks. We propose a word-level text adversarial example generation method. Firstly,design the importance calculation function of the word,then use the classification probability to find the best synonymous replacement word for the word,and then combine the two to determine the replacement order,and finally generate an adversarial sample that is close to the original sample according to the replacement order. The experiments of convolutional neural network,long short-term memory network and bidirectional long short-term memory network models on natural language processing tasks show that the generated adversarial samples reduce the classification accuracy and perturbation rate of the model,and the robustness of the model is improved after adversarial training.

Key words:natural language processing;text adversarial examples;text classification;deep learning;word-level

0 引 言

在大数据的时代和人工智能不断突破新进展、新理论的背景下,深度学习(Deep Learning)[1]早已被广泛应用于计算机视觉[2]、语音识别[3]和自然语言处理(Natural Language Processing,NLP)[4]等热门领域,并且取得了令人瞩目的成就。然而研究表明,深度学习模型容易遭到对抗样本的破坏,这引起了人们对其应用程序中重大安全问题的关注。

对抗样本首先是在图像领域被发现的,而后研究人员在NLP任务中(例如虚假新闻检测、情感分析、文本分类等)也发现了对抗样本。与图像对抗攻击[5]类似,文本对抗攻击[6]的全称为文本对抗样本的生成过程,是指对原始输入中的文本添加微小的且难以察觉的扰动。这种被扰动后的文本依旧会使人类观察者正确分类,却导致了目标模型分类错误。除了达到愚弄目标模型的目的,一个有效的对抗样本还应该满足效能保持的要求。效能保持意味着对抗样本与原始样本相比应在语义上保持不变且语法上保持正确。

虽然文本对抗样本是由图像领域发展而来,但经实验证实图像领域的算法基本上不可直接运用到文本领域,因为图像数据和文本数据有着本质差别。具体来说,图像是像素集合的表示,是连续的;文本是符号化的表示,是离散的。对于人类而言,图像中像素的微小变化不会被感知,表达的含义也没有改变,但对于文

收稿日期:2022-05-31 **修回日期**:2022-09-30
基金项目:安徽省自然科学基金(2008085MF220)
作者简介:张 影(1996-),女,硕士研究生,通信作者,研究方向为网络与信息安全;方贤进,博士,教授,研究方向为机器学习、网络与信息安全、智能计算;杨高明,博士,教授,研究方向为机器学习、隐私保护、生成对抗网络。

本的变化可轻易地察觉。在过去十几年的研究工作中,很多学者提出了大量优秀的对抗文本生成方法,范围从字符级翻转[7]扩展到句子级转述[8],也都取得了良好的效果。相比之下,以单词替换为主的词级生成方法在对抗样本的流畅性和有效性等方面表现的更加突出,因此也成为了 NLP 任务的主要手段。虽然作为主流技术,遗憾的是,关键词的查找和单词排序机制等方面还没有达到理想状态,很难生成质量较高且有效的对抗样本。

1　相关工作

迄今为止,针对 NLP 任务中对抗样本攻击的研究已有了相应重大进展,以下简要介绍此研究的相关工作。

关于字符级的生成方法,Gao 等人[9]设计了特殊的评分函数判断影响分类类别的关键词,并对前 K 个关键词进行随机插入、删除、替换等操作以扰动原始样本。由于扰动是随机的,生成正确单词的概率较低且对抗样本的可用性不高。在文献[10]中细化了 Gao 等人[9]设计的评分函数,在情感分析数据集上验证了改进的有效性。Ebrahimi 等人[11]通过使用热输入向量的梯度在输入中操纵字符级的插入、交换和删除以构建对抗样本。然而字符级别的扰动通常只会改变字符,这会导致语法错误和文本阅读不流畅。关于句子级的生成方法,Jia 等人[12]通过在原始样本末端插入不相关的句子来生成对抗样本;Iyyer 等人[13]利用一种回译数据的神经转述模型将原始句子进行复述,经复述输出的句子作为对抗样本。由于句子级别的扰动颗粒较大,对抗样本与原始样本之间的差别也较大。

相对而言,单词级的扰动在对抗样本的质量及攻击成功率方面表现的更加优异。Papernot 等人[14]随机地替换输入样本中的单词,这种替换方式无疑会破坏原始样本的含义和语法的正确性。Alzantot 等人[15]利用遗传算法(Genetic Algorithm,GA)中的交叉和变异操作生成扰动,减少了对抗样本的替换词数。随后,Wang 等人[16]改进了 GA,它在可以随机剪切单词的基础之上,还可以随机剪切文本片段,攻击效果也有了一定程度的提升。Ren 等人[17]提出了一种基于概率加权词显著性(Probability Weighted Word Saliency,PWWS)的方法,利用同义词替换原始词构造了质量良好的对抗样本,但是生成效率十分低下。Jin 等人[18]提出的 TextFooler 生成算法也利用了同义词作替换词,并且采用了重要单词替换选择机制。虽比 PWWS 的生成效率有所提高,但仍不理想。Wang 等人[19]借助图像领域的梯度攻击提出了快速梯度投影法(Fast Gradient Projection Method,FGPM)应用在 NLP 任务中,虽然对抗样本能够使目标模型判断错误,但这种直接使用在文本上的方法造成对抗样本可读性差。

尽管单词级生成方法更为有效,但是就目前的研究进展来看,在对抗样本的分类准确率和单词扰动百分比方面还有很大的进步空间。针对此现象,该文提出了一种单词级的对抗样本生成方法,对生成过程进行优化以生成质量较高的对抗样本,并通过实验证明了该方法的有效性。

2　研究方法

2.1　问题定义

文本对抗样本是对原始输入文本进行小的修改而形成的,可以改变文本分类器的判断。给定一个包括一切可能的输入文本的特征空间 $X = \{x_1, x_2, \cdots, x_N\}$ 和输出空间 $Y = \{y_1, y_2, \cdots, y_K\}$,原始输入文本可以表示为 $x = [w_1, w_2, \cdots, w_i, \cdots]$。其中 x_N 表示第 N 个样本,y_K 表示样本对应的类别标签,w_i 表示样本中第 i 个单词。一个预训练过的分类器 $f: X \to Y$,它把输入的文本空间 X 映射到标签空间 Y,将原始样本 x 分类到正确标签。对于输入文本 $x \in X$,通过添加不可感知的扰动 Δx 生成对抗样本。一个成功的对抗样本 x_{adv} 应该符合以下约束条件:

$$f(x) \neq f(x_{\text{adv}}) \tag{1}$$

其中,$x_{\text{adv}} = x + \Delta x$,扰动 Δx 要求足够小且有效。

2.2　生成文本对抗样本

针对文本对抗样本的约束条件,提出的基于单词重要性联合分类概率生成对抗样本的方法能够满足上述约束。

如图 1 所示,该方法主要分为三个模块:文本表示与预处理模块、生成样本扰动模块和目标模型预测标签模块。接下来将详细介绍这三个模块,完成整个生成过程。

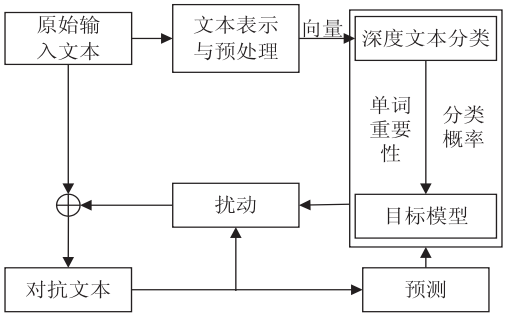

图 1　生成对抗样本整体流程

2.2.1　文本表示与预处理模块

在 NLP 中,将“自然语言”转化为“符号语言”是其本质和核心,所以文本预处理的重要性不言而喻。在此模块中,首先,采用 NLTK[20]自然语言处理工具

包对原始输入文本进行分词处理。由于英文文本的独特性,可直接根据单词间的空格分割开。因此对于每一个样本 x ,其分词结果为 $x=[w_1,\cdots,w_{i-1},w_i,\cdots]$ 。其次,利用 NLTK 的词性标注器对单词进行词性标注。最后,运用预先训练的 100 维 GloVe 词嵌入[21]将文本转化为向量。该过程如图 2 所示。

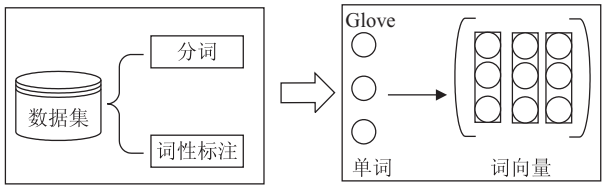

图 2 文本表示与预处理过程

2.2.2 生成样本扰动模块

对抗样本质量的好坏与添加的扰动密不可分。为了生成质量良好的对抗样本,在此模块中主要是生成扰动并对添加的扰动进行约束,利用决策机制在每一步骤中都采用最佳选择。在这项工作中,采用的是基于单词重要性分数联合分类概率选择最佳同义替换词和确定替换顺序。首先,计算单词的重要性分数并过滤停用词,再为其建立同义词集合,根据分类概率从中选择最佳同义词;而替换顺序由单词的重要性分数和最佳同义词的替换效果共同决定。因此,解决问题的关键在于选择最佳候选同义词和确定替换词的顺序。图 3 为生成样本扰动示意图。

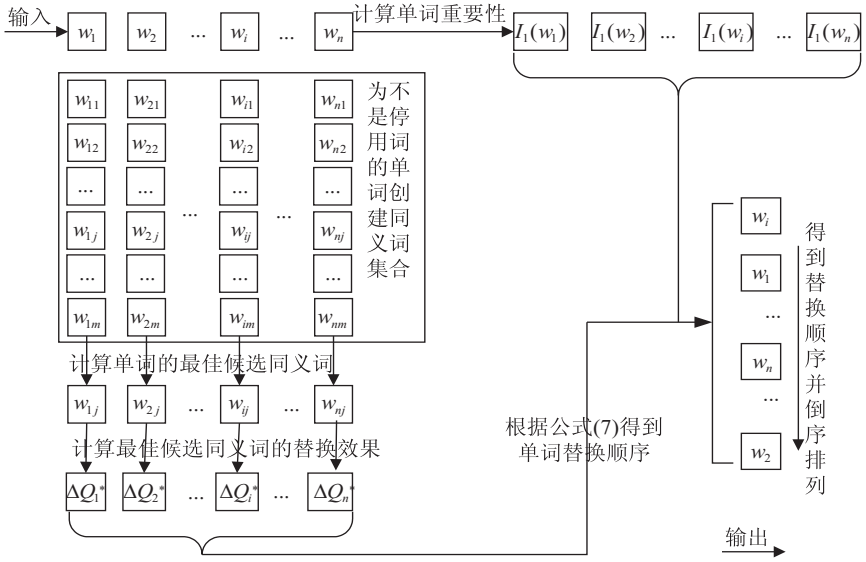

图 3 生成样本扰动示意图

(1)计算单词的重要性得分。

在 NLP 的文本分类任务中,不同的单词对分类结果产生不同程度的影响,需要对文本中的单词进行重要性分数计算。将输入文本删除单词 w_i 后的文本表示为 $x_{\setminus w_i}=[w_1,\cdots,w_{i-1},w_{i+1},\cdots]$, $o_y(x)$ 表示目标分类模型为正确标签 y 输出的逻辑值,并使用 $f(\cdot)$ 表示标签 y 的预测分数。单词重要性分数的计算方式如公式(2)所示:

$$I_1(w_i)=\begin{cases}o_y(x)-o_y(x_{\setminus w_i})\\ 若 f(x)=f(x_{\setminus w_i})=y\\ o_y(x)-o_y(x_{\setminus w_i})+o_{\bar{y}}(x)-o_{\bar{y}}(x_{\setminus w_i})\\ 若 f(x)=y, f(x_{\setminus w_i})=\bar{y}, 且\ y\neq\bar{y}\end{cases}\tag{2}$$

根据公式(2)可看出,若删除单词 w_i 前后分类结果不变,则 $I_1(w_i)$ 为分类模型预测的差值;若删除 w_i 前后分类结果改变,则 $I_1(w_i)$ 为删除 w_i 前后文本被预测为不同类别的差值之和。

一般来说,一段文本语句往往含有不必要的噪声和特征,并不能够对分类器的判定结果起到重要作用,且在不同的文本中,同一单词起到的作用也会有所不同。因此,根据文本特点创建不同的停用词集,过滤“the”、“in”等没有实际意义的单词,可以达到减轻生成扰动的负担的目的。

(2)寻找同义候选词。

为了保证生成对抗样本的单词正确性和语义相似度,即添加扰动后的样本能够使人类尽量无法感知,从而不影响人类的阅读和理解,该文采用与单词的同义词进行替换方式产生扰动,因此对于本方法而言,扰动的最初状态是对输入单词所查找的同义词。

具体来说,首先,使用 WordNet 数据语料库[22]为单词 w_i 构建一个同义词集合 $\mathbb{L}_i$,每个同义词 $w_{ij}\in\mathbb{L}_i$ 都是单词 w_i 的同义候选词。其次,从集合 $\mathbb{L}_i$ 中选择一个同义词 w_{ij} 替换 w_i 得到样本 $x_i^{'}$,可表示为 $x_i^{'}=[w_1,\cdots,w_{i-1},w_{ij},\cdots]$, w_{ij} 表示 $\mathbb{L}_i$ 中第 j 个同义词。而每一个同义词替换后,样本 $x_i^{'}$ 的预测类别会有所不同。因此,通过采用替换同义词 w_{ij} 前后分类概率的差值的方法比较它们造成的替换效果以确定最佳替换词。该替换效果的计算如公式(3)所示:

$$I_2(w_i) = o_y(x) - o_y(x_i') \tag{3}$$

根据经验可知,使分类概率差值达到最大的同义词得到的替换效果最好,因此, w_i 对应的最佳替换单词 w_i^* 为:

$$w_i^* = \underset{w_{ij} \in L_i}{\operatorname{argmax}} \{ o_y(x) - o_y(x_i') \} \tag{4}$$

与此同时,得到新的扰动文本:

$$x_i^* = [w_1, \cdots, w_i^*, \cdots] \tag{5}$$

因而,单词 w_i 对应的最佳同义替换词 w_i^* 达到的最佳替换效果如公式(6)所示:

$$\Delta Q_i^* = o_y(x) - o_y(x_i^*) \tag{6}$$

(3)确定最佳替换顺序。

在文本对抗样本中,另一个影响其质量的重大因素就是单词的替换顺序,因为它与样本被扰动的比例息息相关。而单词的替换顺序排列不是单方面起作用,它不仅与重要单词有关,还包括与之相匹配的最佳同义词 w_i^* 。故文中单词的替换顺序由单词 w_i 的重要性得分和其最佳替换同义词产生的替换效果共同决定。也即对公式(2)中的第 i 个值以及公式(6)中替换词进行打分,如公式(7)所示:

$$\text{Score}(x, x_i^*, w_i) = \Phi(I_1(w_i))_i \cdot \Delta Q_i^* \tag{7}$$

根据打分结果对单词倒序排序以获得最佳替换顺序 w_{order} 。为确保生成质量更佳的对抗样本,再次为扰动添加约束。因为每个文本具有不同长度,所以设置替换上限为该文本单词数量的 6% 。

2.2.3　目标模型预测标签模块

目标模型预测标签模块是该算法的最后一步,旨在判别对抗样本的标签,它的作用是验证生成的对抗样本是否有效。此模块的具体流程如图 4 所示。先将生成的扰动根据顺序 w_{order} 添加到原始样本生成扰动样本,再输入到目标分类模型中得到类别标签,由标签结果决定是否进行更多的替换。若标签改变即成功,否则继续进行替换直到文本分类的结果改变,或者替换单词数量达到上限即失败。

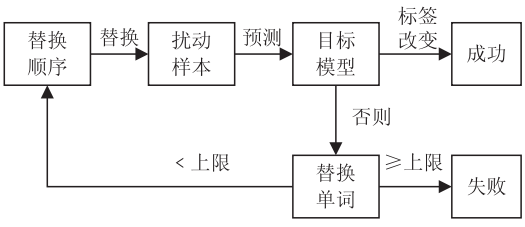

图 4　目标模型预测标签流程

采用的主要算法过程如下所示:

算法:对抗样本生成过程。

输入:原始样本 $x = [w_1, \cdots, w_{i-1}, w_i, \cdots]$,标签 y ,文本分类模型 f ,替换上限比例 6% top;

输出:对抗样本 x_{adv} 。

1:初始化: $x_{\text{adv}} \leftarrow x$

2:for each w_i in x do:

3:根据等式(2)计算单词重要性分数 $I_1(w_i)$;

4:end for

5:建立停用词集 stop_words set;

6:for word w_i in x do:

7:　if w_i not in stop_words set:

8:　为单词 w_i 建立同义词集合 L_i ;

9:end for

10:for word w_i in x do:

11:　for candidate word w_{ij} in L_i do:

12:　根据等式(3)计算同义候选词的替换效果 $I_2(w_i)$

13:　end for

14:　根据等式(4)得到单词 w_i 的最佳替换词 w_i^*

15:end for

16: $w_{\text{order}} \leftarrow$根据 $\text{Score}(x, x_i^*, w_i)$ 计算单词的替换顺序并倒序排列;

17:for word w_i in w_{order} do:

18:　$x_i' \leftarrow$将单词 w_i 用最佳替换词替换;

19:　$x_{\text{adv}} = x_i'$;

20:　if $f(x) \neq f(x_{\text{adv}})$:

21:　　return x_{adv}

22:　else if 单词替换数量达到上限 top:

23:　　return none

24:end for.

在上述算法中,2-4 行为计算单词的重要性分数,第 5 行为建立停用词集,6-9 行为单词建立同义候选词集合 $\mathbb{L}_i$,10-15 行为计算同义词的替换效果,确定最佳替换词 w_i^* ,第 16 行为根据 $\text{Score}(x, x_i^*, w_i)$ 确定单词的替换排序,17-24 行为根据 w_{order} 中的单词依次替换,直到预测标签改变返回“成功”输出对抗样本或替换数量达到上限返回“失败”。

3　实验设置与结果分析

计算机将文本数据信息映射到给定的某一类别或某几类别标签的过程称为文本分类(Text Classification)。它被广泛应用在新闻主题分类、情感分析、舆情分析及邮件过滤等场景中。在这一节中,分析了所提出的方法在文本分类任务上的性能。

3.1　数据集与目标分类模型

为了验证生成方法的有效性,在三个流行的数据集上对不同的深度神经网络模型进行分类预测,分别为 IMDB 数据集、AG´s News 数据集和 Yahoo! Answers 数据集。

IMDB 数据集[23]:该数据集是用于情感分类任务,数据集中的每个样本都是一个电影评论,类别标签为积极或消极。

AG´s News 数据集[24]:该数据集是用于新闻文章分类,由世界新闻、体育新闻、商业新闻和科学新闻四

个类别组成。

Yahoo! Answers 数据集[25]:该数据集由十大类别主题组成,包含 1 400 000 个训练样本和 5 000 个测试样本,平均分布在不同的类别上。

目标模型选用卷积神经网络[26](Convolutional Neural Networks, CNN)、长短时记忆网络[27](Long Short-Term Memory, LSTM)和双向长短时记忆网络[28](Bi-directional Long Short-Term Memory, Bi-LSTM),并选用 Random、FGPM[19]及 PWWS[17]作为实验的对比方法评估文中方法的性能,且每组实验都是从数据集中随机抽取 2 000 个干净样本进行扰动。其中 Random 是指随机地选择样本中的单词进行替换,并没有预先计算替换顺序。

3.2 评估指标

为了评估对抗样本的质量,在实验中设置了两种评估指标,分别为分类正确率(Classification Accuracy, CA)和扰动率(Perturbation Rate, PR)。分类正确率是指对抗样本被模型分类准确的比例。它是衡量生成方法成功与否的核心指标,也表示着误导模型的能力。值越小,结果越有效。扰动率是指文本被扰动的比例,也是评估对抗样本质量的重要因素。一般而言,越少的扰动表示着语义一致性越高。两者的计算公式分别如下:

$$CA = success\_count / sum\_count \tag{8}$$

$$PR = substitute\_count / len(doc) \tag{9}$$

其中, success_count 表示对抗样本标签改变的个数, sum_count 表示输入样本的总个数, substitute_count 为被替换单词的数量; len(doc) 为样本的长度。

3.3 评估结果分析

表 1 和表 2 分别展示了文中方法与其他方法在 AG's News 数据集和 Yahoo! Answers 数据集上三个模型的分类准确率对比。由对抗样本误导分类标签的性质可知,对抗样本的分类准确率越小越好。在每组实验中,文中的生成方法都能使得模型达到较低的分类准确率,证明它可以最大程度地欺骗模型并显著地降低性能。实验结果还表明,相比于其他分类模型,LSTM 模型使得对抗样本的分类准确率下降的最为显著。

表 1 AG's News 数据集生成对抗样本的分类准确率(CA)

目标模型	无扰动	Random	FGPM	PWWS	文中方法
CNN	0.908 9	0.741 3	0.375 0	0.396 0	0.352 0
LSTM	0.926 0	0.724 4	0.310 0	0.300 0	0.190 0
Bi-LSTM	0.903 8	0.704 3	0.320 0	0.150 0	0.120 0

表 2 Yahoo! Answers 数据集生成对抗样本的分类准确率(CA)

目标模型	无扰动	Random	FGPM	PWWS	文中方法
CNN	0.892 1	0.820 9	0.060 0	0.102 0	0.086 0
LSTM	0.751 0	0.691 0	0.170 0	0.125 0	0.084 0
Bi-LSTM	0.870 8	0.796 6	0.105 0	0.110 0	0.096 0

此外如表 3 中第三列所示,在 IMDB 数据集上再次验证了该方法可降低对抗样本被模型分类正确的概率。如前所述,扰动率越小,对抗样本与原始样本越接近。在第四列中可清晰地观察到扰动率保持在较小值,亦即替换的单词数量较少,可获得较强的隐蔽性,从而使得对抗样本在语义和语法上与原始文本保持较大的一致性。

表 3 在 IMDB 数据集生成对抗样本的分类准确率(CA)和扰动率(PR)

生成方法	目标模型	CA	PR
FGPM	CNN	0.083 0	0.051 0
	LSTM	0.184 3	0.071 1
	Bi-LSTM	0.053 3	0.040 1
PWWS	CNN	0.103 0	0.059 7
	LSTM	0.159 1	0.065 0
	Bi-LSTM	0.045 0	0.033 8
文中方法	CNN	0.060 0	0.047 8
	LSTM	0.136 0	0.059 1
	Bi-LSTM	0.011 0	0.032 1

综上所述,所提出的方法能够在降低分类模型准确率的同时减小样本的扰动比例,使得生成的对抗样本与原始样本保持较高的一致性,不容易被人类察觉。

3.4 对抗训练及案例分析

对抗训练的目的是为了提高模型的鲁棒性。为了验证模型是否提高了正视对抗样本的能力,将 IMDB 数据集生成的对抗样本加入到其干净样本中,构造一个新的数据集来重新训练 CNN 模型。结果如图 5 所示,随着对抗样本数量的增加,目标模型能够更好地拟合这些数据,模型的分类准确率逐步上升。换句话说,

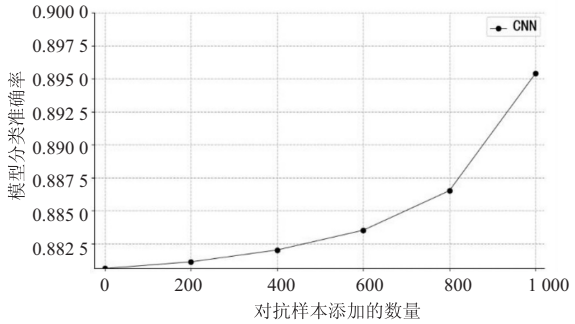

图 5 对抗训练

对抗样本被经过对抗训练后的模型分类到错误类别的概率在减小,模型的脆弱性得到了保护,也就有了更强的鲁棒性。

表 4 展示了在 IMDB 数据集上针对 CNN 模型生成文本对抗样本的示例。从表中可以看出,在该文本中仅替换 1 个单词,就实现了将原始标签“积极”错误地判别为“消极”。

表 4　IMDB 数据集上原始样本和对抗样本示例

原始样本标签:积极	对抗样本标签:消极
A quite easy to watch tale of 2 thieves, with that love/hate type relationship between them. ChrisopherWalken stars and is verydear as the silent rogue with a scam bigger than he´s letting on.	A quite easy to watch tale of 2 thieves, with that love/hate type relationship between them. ChrisopherWalken stars and is verygood as the silent rogue with a scam bigger than he´s letting on.

4　结束语

在 NLP 文本分类任务中,针对目标分类模型的脆弱性,提出了一种单词级的对抗样本生成方法。该方法在单词重要性和分类概率的共同作用下生成微小的扰动。在三个文本分类数据集上的实验表明,对抗样本在保持较低的分类正确率的同时具有较低的扰动率,亦即对抗样本质量有所提高。进一步的实验表明,模型在对抗训练后提高了自身的鲁棒性。

参考文献:

[1] LECUN Y,BENGIO Y,HINTON G. Deep learning[J]. Nature,2015,521(7553):436-444.

[2] CHAI J,ZENG H,LI A,et al. Deep learning in computer vision:a critical review of emerging techniques and application scenarios[J]. Machine Learning with Applications,2021,6:100134.

[3] 更藏措毛,黄鹤鸣. 双向循环神经网络在语音识别中的应用[J]. 计算机与现代化,2019(10):1-6.

[4] 罗　枭. 基于深度学习的自然语言处理研究综述[J]. 智能计算机与应用,2020,10(4):133-137.

[5] 陈梦轩,张振永,纪守领,等. 图像对抗样本研究综述[J]. 计算机科学,2022,49(2):92-106.

[6] 杜小虎,吴宏明,易子博,等. 文本对抗样本攻击与防御技术综述[J]. 中文信息学报,2021,35(8):1-15.

[7] HE X,LYU L,SUN L,et al. Model extraction and adversarial transferability,Your BERT is Vulnerable! [C]//Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics:human language technologies. Mexico:Association for Computational Linguistics,2021:2006-2012.

[8] ZHANG Y,BALDRIDGE J,HE L. PAWS:paraphrase adversaries from word scrambling[C]//Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics:human language technologies,volume 1(long and short papers). Minneapolis:Association for Computational Linguistics,2019:1298-1308.

[9] GAO J,LANCHANTIN J,SOFFA M L,et al. Black-box generation of adversarial text sequences to evade deep learning classifiers[C]//2018 IEEE security and privacy workshops (SPW). San Francisco:IEEE,2018:50-56.

[10] LI J,JI S,DU T,et al. TextBugger:generating adversarial text against real-world applications[C]//Proceedings of the 26th annual network and distributed system security symposium. San Diego:NDSS,2019:24-27.

[11] EBRAHIMI J,RAO A,LOWD D,et al. HotFlip:white-box adversarial examples for text classification[C]//Proceedings of the 56th annual meeting of the association for computational linguistics (volume 2:short papers). Melbourne:Association for Computational Linguistics,2018:31-36.

[12] JIA R,LIANG P. Adversarial examples for evaluating reading comprehension systems[C]//Proceedings of the 2017 conference on empirical methods in natural language processing. Copenhagen:Association for Computational Linguistics,2017:2021-2031.

[13] IYYER M,WIETING J,GIMPE K,et al. Adversarial example generation with syntactically controlled paraphrase networks[C]//Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics:human language technologies,volume 1 (long papers). New Orleans:Association for Computational Linguistics,2018:1875-1885.

[14] PAPERNOT N,MCDANIEL P,SWAMI A,et al. Crafting adversarial input sequences for recurrent neural networks [C]//MILCOM 2016-2016 IEEE military communications conference. Baltimore:IEEE,2016:49-54.

[15] ALZANTOT M,SHARMA Y,ELGOHARY A,et al. Generating natural language adversarial examples[C]//Proceedings of the 2018 conference on empirical methods in natural language processing. Brussels:Association for Computational Linguistics,2018:2890-2896.

[16] WANG X,JIN H,HE K. Natural language adversarial attacks and defenses in word level[J]. arXiv:1909.06723,2019.

[17] 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:Association for Computational Linguistics,2019:1085-1097.

[18] 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]//Proceedings of the AAAI conference on artificial intelligence. New York:AAAI,2020:8018-8025.

[19] WANG X, YANG Y, DENG Y, et al. Adversarial training with fast gradient projection method against synonym substitution based text attacks[C]//Proceedings of the AAAI conference on artificial intelligence. Vancouver: AAAI, 2021: 13997-14005.

[20] EMMERY C, KÁDÁR Á, CHRUPAŁA G. Adversarial stylometry in the wild:transferable lexical substitution attacks on author profiling[C]//Proceedings of the 16th conference of the european chapter of the association for computational linguistics:main volume. Kyiv:Association for Computational Linguistics,2021:2388-2402.

[21] 吉久明,施陈炜,李 楠,等. 基于 GloVe 词向量的“技术——应用”发现研究[J]. 现代情报,2019,39(4):13-22.

[22] MOZES M, BARTOLO M, STENETORP P, et al. Contrasting human-and machine-generated word-level adversarial examples for text classification[C]//Proceedings of the 2021 conference on empirical methods in natural language processing. Online and Punta Cana:Association for Computational Linguistics,2021:8258-8270.

[23] MENG Z, WATTENHOFER R. A geometry-inspired attack for generating natural language adversarial examples[C]//Proceedings of the 28th international conference on computational linguistics. Barcelona: International Committee on Computational Linguistics,2020:6679-6689.

[24] LI L, MA R, GUO Q, et al. BERT-ATTACK:adversarial attack against BERT using BERT[C]//Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP). Online:Association for Computational Linguistics,2020:6193-6202.

[25] YANG P, CHEN J, HSIEH C J, et al. Greedy attack and gumbel attack: generating adversarial examples for discrete data[J]. Journal of Machine Learning Research, 2020, 21(43):1-36.

[26] CHEN Y. Convolutional neural network for sentence classification[D]. Waterloo:University of Waterloo,2015.

[27] YOO J Y, MORRIS J, LIFLAND E, et al. Searching for a search method: benchmarking search algorithms for generating NLP adversarial examples[C]//Proceedings of the third blackbox NLP workshop on analyzing and interpreting neural networks for NLP. Online: Association for Computational Linguistics,2020:323-332.

[28] MALIK V, BHAT A, MODI A. Adv-OLM:generating textual adversaries via OLM[C]//Proceedings of the 16th conference of the european chapter of the association for computational linguistics:main volume. Kyiv:Association for Computational Linguistics,2021:841-849.

(上接第 97 页)

Communication and Computing,2018,29:103-112.

[8] GRIGORIEV D. Complexity of solving tropical linear systems[J]. Computational Complexity,2013,22:71-88.

[9] MAZE G, MONICO C, ROSENTHAL J. Public key cryptography based on semigroup actions[J]. Advances of Mathematics of Communications,2007,1(4):489-507.

[10] STEINWANDT R, CORONA A. Cryptanalysis of a 2-party key establishment based on a semigroup action problem[J]. Advances in Mathematics of Communications,2011,5(1):87-92.

[11] ATANI R E, ATANI S E, MIRZAKUCHAKI S. Public key cryptography based on semimodules over quotient semirings [J]. Int. Mathematical Forum, 2007, 2(49-52):2561-2570.

[12] DURCHEVA M. Public key cryptosystem based on two sided action of different exotic semirings[J]. Journal of Math and System Science,2014,4:6-13.

[13] AHMED K, PAL S, MOHAN R. A review of the tropical approach in cryptography[J]. Cryptologia,2021,32(2):1-25.

[14] GRIGORIEV D, SHPILRAIN V. Tropical cryptography[J]. Communications in Algebra,2014,42(6):2624-2632.

[15] KOTOV M, USHAKOV A. Analysis of a key exchange protocol based on tropical matrix algebra[J]. Journal of Mathematical Cryptology,2018,12(3):137-141.

[16] GRIGORIEV D, SHPILRAIN V. Tropical cryptography II: extensions by homomorphisms[J]. Communications in Algebra,2019,47(10):4224-4229.