Adversarial Examples Generation Method for Chinese Text Classification*

GONG Yan[1], ZHANG Xiaolin[1*], LIU Yuefeng[1], LIU Lixin[1,2], XU Li[3]
(1.School of Information Engineering, Inner Mongolia University of Science and Technology, Baotou Inner Mongolia 014010, China;
2.School of Information, Renmin University of China, Beijing 100872, China;
3.Department of Computer Science and Technology, Baotou Medical College, Baotou Inner Mongolia 014010, China)

Abstract: Aiming at the robustness of deep neural networks, a black-box adversarial example generation method for Chinese text classification, i.e., WordBeguiler is proposed, which combines Chinese character shape and phonetic features to construct the adversarial search space, finds important words or phrases that affect the classification results by designing new perturbation localization methods, and determines modification strategies to generate adversarial examples according to the method of probability weights selection. Two mainstream models, convolutional neural network (CNN) and bidirectional long short-term memory network (BiLSTM), are used to verify their effectiveness and transferability on different classification data sets. The experimental results show that WordBeguiler has the advantage of a higher attack success rate and lower perturbation rate compared to other attack methods while retaining the original semantics and ensuring the grammatical correctness to a certain extent, and can be effectively transferred to the BERT model.

Key words: Chinese text classification; adversarial examples; deep neural network; Chinese character characteristics; black box
EEACC: 6140 **doi**: 10.3969/j.issn.1005-9490.2023.05.029

面向中文文本分类的对抗样本生成方法*

弓 燕[1], 张晓琳[1*], 刘月峰[1], 刘立新[1,2], 徐 立[3]
(1.内蒙古科技大学信息工程学院, 内蒙古 包头 014010; 2.中国人民大学信息学院, 北京 100872;
3.包头医学院计算机科学与技术系, 内蒙古 包头 014010)

摘 要: 针对深度神经网络鲁棒性问题, 提出了一种面向中文文本分类的黑盒对抗样本生成方法 WordBeguiler。该方法结合汉字的字形、字音特征构建对抗搜索空间, 设计了新的扰动定位方式寻找影响分类结果的重要字或词组, 并根据概率权重选取的方法确定修改策略生成对抗样本。使用两个主流的模型卷积神经网络(CNN)和双向长短期记忆网络(BiLSTM)在不同分类数据集上验证其有效性和可转移性。实验结果表明, 与其他攻击方法相比, WordBeguiler 具有攻击成功率高、扰动率低的优势, 同时既保留了原始语义也一定程度上保证了语法正确性, 并且可以有效地转移到 BERT 模型中。

关键词: 中文文本分类; 对抗样本; 深度神经网络; 汉字特征; 黑盒
中图分类号: TP309 **文献标识码**: A **文章编号**: 1005-9490(2023)05-1349-08

深度神经网络的应用为人工智能领域取得了突破性的进展, 如图像分类[1]、语音识别[2]、自然语言处理[3]等。但是, 由于神经网络自身存在局部线性等特点, 使其极易受到安全性威胁。Szegedy 等[4]首次提出在图像分类任务中对图像添加难以察觉的扰动可诱导模型错误分类。此后, 针对计算机视觉领域对抗样本生成方法的研究相继展开, 例如经典的对抗样本生成算法有基于快速梯度投影方法(Fast Gradient Sign Method, FGSM)[5], 投影梯度下降算法(Project Gradient Descent, PGD)[6], 基于雅可比矩阵的显著性图攻击(Jacobian-based Saliency Map Attack, JSMA)[7]等。

自然语言处理包括情感分析[8]、文本分类[9]、机器翻译[10]以及问答系统[11]等任务, 同样也存在着安全性问题。与图像不同的是, 文本具有离散性, 用梯度方法为输入数据添加扰动易察觉, 而且会造成语法错误、语义不连贯等问题, 使用的距离度量方式也不一致。因此, 在图像领域中对抗样本的生成方法难以直接应用于文本。目前, 文本对抗样本的生成方法面临着三个挑战。第一, 如何保证对抗样本与原始样本的语义一致性; 第二, 如何保持对抗样本的语法正确性; 第三, 如何提高对抗扰动的不可被察觉性。

在文本领域, Papernot 等[12]通过 FGSM 找到输入序列中梯度最大的位置, 即重要单词的位置, 使用插入、修改、和删除三种扰动策略生成对抗样本, 从而误导字符级模型和词级模型, 但该方法是在已知模型参数的前提下实现的, 并不适用于黑盒场景[13-14]; Gao 等[15]提出的 DeepWordBug 攻击算法, 通过设计重要性评分函数找到重要单词, 采用插入、替换、交换前后

项目来源: 国家自然科学基金项目(61562065); 内蒙古自治区自然科学基金项目(2019MS06001; 2019MS06036)
收稿日期: 2022-04-26 **修改日期**: 2022-06-20

字母的位置以及删除操作生成对抗样本欺骗分类器；Garg 等[16]提出的基于 BERT 生成的对抗样本方法(BERT-based Adversarial Examples，BAE)算法，是第一个使用语言模型生成对抗样本的方法，使用评分函数找到重要单词的位置，然后利用 Bert 语言模型对其掩码与预测，选择具有对抗性且与原句相似度最大的单词作为替换词。以上方法均在英文文本中实现，由于英文与中文在字符的种类、字的长度以及音形特点上有着明显的差异，因此面向英文文本生成对抗样本的方法不能直接应用到中文文本。

目前，面向中文文本对抗样本生成方法的研究处于起步阶段，王文琦等[17]提出 WordHanding 攻击算法，能够在黑盒条件下攻击中文文本的情感倾向性系统，借助长短期记忆网络(Long Short-Term Memory Network，LSTM)预训练模型判别分类倾向，利用词语重要性函数计算输入序列中字词的重要程度，采用同音词替换生成对抗样本，达到了较好的攻击效果且扰动字数少，但替换方式单一；仝鑫等[18]提出一种在词级黑盒条件下生成对抗样本的方法 CWordAttacker，采用定向词评分删除机制计算字词的重要性，通过繁体字替换、拼音改写、词组拆解以及次序扰动四种攻击策略生成对抗样本，但是扰动率较大，未能保持较高的语义相似度。

基于前人的研究工作，提出了一种面向中文文本分类的黑盒对抗样本生成方法 WordBeguiler，通过分析汉字的字形、字音特征构建对抗搜索空间，设计一种新的扰动定位方式计算影响模型分类的重要字词，并根据概率权重选取的方法确定修改策略生成对抗样本，有效地实现了在多场景分类任务下面向中文文本的对抗攻击。

本文主要贡献有：

①提出了一种面向中文文本分类的黑盒对抗样本生成方法 WordBeguiler，在无需了解目标模型内部参数的情况下，只对输入文本添加微小的扰动即可生成对抗样本，适用于情感分析、垃圾邮件检测以及新闻分类等任务。

②设计了结合汉字字形、字音特征构建对抗搜索空间的方法，有效地提高了对抗样本的语法正确性，增强不可察觉性。

③设计了一种新的扰动定位方式，能够有效地找出不同分类任务下影响模型决策的关键字词，降低修改率的同时提高了攻击成功率。

④在三个公开数据集上进行实验，通过攻击卷积神经网络(Convolutional Neural Network，CNN)和双向长短期记忆(Bi-directional Long Short-Term Memory，BiLSTM)模型，使准确率降低了百分之五十左右，并且具有较高的可转移性。

1 背景知识

1.1 卷积神经网络 CNN

CNN[19]是一种通过卷积计算的神经网络，由输入层、卷积层、池化层、全连接层以及 softmax 层组成。其中，卷积层是 CNN 最重要的部分，使用卷积核对窗口中的元素依次进行点乘、相加运算。池化层对输入数据不同位置特征进行采样，在保证收集到重要特征的条件下压缩参数量，防止过拟合。经过多轮的卷积和池化操作后，原信息被抽象成了信息含量更高的特征，提取到的特征传送到全连接层进行整合，最后经过 softmax 层，将输出映射到 0~1 之间，表示分类概率。

在短文本分类任务中，CNN 可以保留文本的空间信息，同时捕获到局部信息和顺序信息，可应用于情感分析、垃圾邮件检测等任务。

1.2 BiLSTM

BiLSTM[20]是 LSTM[21]的一种扩展形式，采用双向网络结构充分地提取上下文信息作为分类依据，有效地提高了文本分类的准确率。BiLSTM 首先将 embedding 后的词向量输入网络，然后经过前向 LSTM 层和后向 LSTM 层得到隐向量，并进行拼接作为网络的输出，最后通过全连接层和 softmax 层得到最终的分类结果，网络结构如图 1 所示。

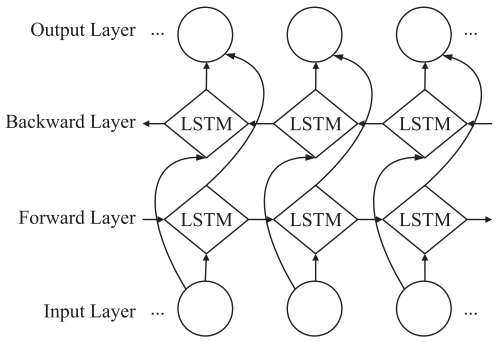


图 1 BiLSTM 网络结构图

1.3 文本对抗样本

对于数据集中的 N 个文本 $X=\{x_1,x_2,\cdots,x_N\}$，其对应的分类标签为 $Y=\{y_1,y_2,\cdots,y_M\}$，F 是分类模型，学习从输入文本 $x\in X$ 到标签 $y\in Y$ 的映射关系 $F:X\to Y$。通过向原始输入文本 x 中添加微小的扰动 Δx 生成对抗样本 x^*，从而迫使深度学习模型 F 错误分类。如式(1)所示：

$$x^*=x+\Delta x, F(x)=y,\quad F(x^*)\neq y \text{ or } F(x^*)=y', y'\neq y \tag{1}$$

式中:$\|\Delta x\|<\delta$,δ 是限制扰动大小的阈值。

同时,一个好的对抗样本 x^* 不仅可以欺骗 F,而且还需满足三个条件,即语义一致性、语法正确性以及不可察觉性,本文提出的对抗样本生成方法很好地满足了这三点。

2　WordBeguiler 算法

2.1　问题定义

给定文本分类模型 F,学习从文本到标签的映射关系 $F:X\rightarrow Y$,将输入文本空间 X 映射到标签空间 Y,对于句子 $x\in X$,其真实标签 $y\in Y$,有效的对抗样本 x^* 应满足:

$$F(x^*)\neq F(x),\text{Cost}(x^*,x)\leqslant\sigma \tag{2}$$

式中:Cost(·)表示文本修改累计频次,σ 表示修改最大阈值。

2.2　总体描述

本文设计的中文文本分类的对抗样本生成算法具体流程如图 2 所示。

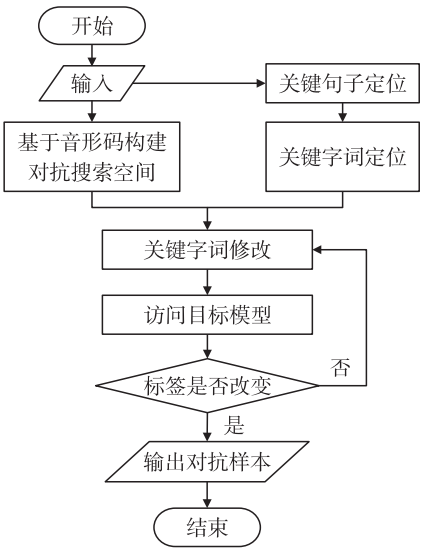


图 2　对抗样本生成流程

具体过程描述如下:

①基于音形码构建对抗搜索空间:结合汉字的字形和字音特征为每个汉字构建对抗搜索空间,且将繁体字、拆字和拼音加入其中。

②扰动定位:首先关键句子定位,将原始文本按照标点符号划分为若干个句子,去除预测标签与文本的真实标签不一致的句子,通过关键句子评分函数计算每个句子的重要性,降序排列。其次关键字词定位,对每一个句子进行分词,过滤掉停止词,按照关键字词评分函数筛选出重要字词。

③关键字词修改:本文共设置了三种修改策略,分别是字词替换、字符插入和次序修改,采用概率权重选取的方法确定修改策略。

将修改后的样本输入到目标模型中,若标签改变则成功生成对抗样本,否则继续返回到步骤③,直至标签改变或者达到最大修改阈值为止。

2.3　基于音形码构建对抗搜索空间

对抗搜索空间的质量决定着对抗样本的质量,因此,文本结合汉字的字形和字音特点构建对抗搜索空间以提高对抗样本质量。首先对字典中的汉字进行音形编码,其次计算每一个字与其他汉字之间的音形码汉明距离,根据汉明距离计算相似度,最后按照相似度降序排列,选取前 K 个加入对抗搜索空间。

2.3.1　音形编码方式

编码包括拼音和字形两部分。

①拼音部分:汉字的拼音结构由声母、韵母以及声调构成,所以对拼音进行编码也应包含这三部分。对于某些汉字拼音中声母与韵母之间存在一个元音的情况,如 nuan 中的 u,miao 中的 i 等,需额外添加一个补码位。

第一部分表示声母,声母有 23 种,用五位二进制数表示。为了在后期计算相似度的时候弱化平翘舌音差异,因此这里 zh 与 z,ch 与 c,sh 与 s 采用同一个编码表示。

第二部分表示韵母,韵母有 24 种,用五位二进制数表示。同理,前后鼻音也采用同一个编码表示。声母编码和韵母编码均采用格雷码形式,可使相邻两个读音相近的编码差异最小。拼音声母编码和韵母编码表示方式如表 1 和表 2 所示。

表 1　拼音声母编码

b=00000	p=00001	m=00011	f=00010
d=00111	t=00101	n=00100	l=01100
g=01111	k=01110	h=01010	
J=01001	q=01000	x=11000	
zh=11011	ch=11010	sh=11110	
z=11011	c=11010	s=11110	r=11111
y=11100	w=10101		

表 2　拼音韵母编码

a=00000	ai=00001	ao=00011	an=00010	a=00000
i=00111	ie=00101	iu=00100	in=01100	i=00111
o=01111	ou=01110	ong=0100	ang=11101	o=01111
e=01001	ei=01000	er=11000	en=11001	e=01001
u=11011	ui=11110	un=11111	ing=01100	u=11011
ü=11100	üe=10100	ün=10101	eng=11001	ü=11100

第三部分是补码,采用与韵母编码相同的方式,若没有补码则表示为五个 0。

第四部分是声调,声调包括四种,用两位二进制数表示即可,00 表示一声,01 表示二声,10 表示三声,11 表示四声。

因此,音码部分二进制数共有 5+5+5+2=17 位。

②字形部分:汉字的字形特征包括结构,四角号码和笔画,因此需对这三个部分进行编码。

第一部分是字形的结构。由于汉字具有 14 种不同的结构,因此用四位二进制数来表示,同样采用格雷码形式使字形相近的结构在编码表示上相近。结构编码如表 3 所示。

表 3 结构编码

结构	编码	结构	编码
独体字	0000	右上包围	0100
左右结构	0001	上三包围	1100
左中右结构	0011	下三包围	1101
上下结构	0010	左三包围	1111
上中下结构	0110	全包围	1110
左上包围	0111	穿插结构	1010
左下包围	0101	品字结构	1011

第二部分为汉字的四角号码,用于描述汉字的形态特征,每个汉字都可用 0 到 9 中的四个数字表示,通过查找四角号码检字法可得到对应的四角号码。

第三部分是汉字的笔画,笔画数用 16 位二进制数表示。即笔画数 z 的汉字编码方式如式(3)所示:

$$\begin{cases} 0\text{xFF} & z>16 \\ 2^z-1 & z\leqslant 16 \end{cases} \tag{3}$$

因此,形码部分二进制位数共有 4+4×4+16=36 位。

2.3.2 相似度计算方法

由于音码部分和形码部分的二进制位数不一致,会对最终的相似度产生不同影响,因此需计算音码部分和形码部分在最终相似度计算中所占的贡献比 b_1 和 b_2,满足 $b_1+b_2=1$,如式(4)和式(5)所示。其中 q_p 表示音码的汉明距离,q_x 表示形码的汉明距离,l_1 和 l_2 分别表示音码和形码的二进制长度。

$$b_1=\frac{e^{\frac{q_p}{l_1}}}{e^{\frac{q_p}{l_1}}+e^{\frac{q_x}{l_2}}} \tag{4}$$

$$b_2=\frac{e^{\frac{q_x}{l_2}}}{e^{\frac{q_p}{l_1}}+e^{\frac{q_x}{l_2}}} \tag{5}$$

计算汉字之间的相似度 S,如式(6)所示:

$$S=1-\frac{q_p}{l_1}\times b_1-\frac{q_x}{l_2}\times b_2 \tag{6}$$

将相似度降序排列,取前 K 个字加入对抗搜索空间。此外,将该汉字的繁体字,拆字以及对应的拼音形式也添加到对抗搜索空间中。

2.4 扰动定位

2.4.1 关键句子定位

当人们表达自己的观点时,大部分的句子都是在陈述事实,而对分类结果产生较大影响的只有少数几个句子。因此,为了提高 WordBeguiler 的效率,本文先定位对最终预测结果贡献最大的关键句子。

首先将输入文本根据标点符号划分为 n 个子句,得到 $x=\{s_1,s_2,s_3,\cdots,s_n\}$,$s_i$ 表示第 i 个句子,其次过滤掉预测标签与文本的真实标签不一致的句子,最后用预测为真实标签 y 的置信度值作为该句的重要性分数 Cs_i,根据分数降序排列。如式(7)所示:

$$Cs_i=F_y(s_i) \tag{7}$$

2.4.2 关键字词定位

通过上一步定位到关键句子后,进行分词处理得到 $s_i=\{w_1,w_2,w_3,\cdots,w_n\}$,发现只有一些关键字词对模型的预测结果有影响,因此,首先过滤掉停止词,其次使用重要性评分函数去定位关键字词,如式(8)所示:

$$I_{w_j}=\begin{cases} F_y(s_i)-F_y(s_{i\setminus w_j}) \\ \quad \text{if } F(s_i)=F(s_{i\setminus w_j})=y \\ (F_y(s_i)-F_y(s_{i\setminus w_j}))+(F_{y'}(s_{i\setminus w_j})-F_{y'}(s_i)) \\ \quad \text{if } F(s_i)=y, F(s_{i\setminus w_j})=y' \text{ and } y\neq y' \end{cases} \tag{8}$$

式中:I_{w_j}作为字词 $w_j\in s_i$ 对于分类结果 $F(s_i)=y$ 的重要性分数,$s_{i\setminus wj}=\{w_1,\cdots,w_{j-1},w_{j+1},\cdots,w_n\}$ 表示删除字词 w_j 的句子,$F_y(\cdot)$表示预测为 y 标签的置信度。

2.5 关键字词修改

为了保证生成的对抗样本与原样本语义相近,即在误导深度神经网络模型做出错误预测的情况下有较好的可读性,WordBeguiler 算法采用以下三种修改策略:①字词替换:利用基于音形码构建的对抗搜索空间进行替换;②字符插入:随机插入不常见的特殊符号。在此建立了一个符号集,包括标点符号、罗马字符等;③次序修改:交换相邻两个字的顺序。据心理学研究表明,交换汉字顺序不会影响人类对文本的阅读理解。对于上述修改策略,采用概率权重选取的方法确定。

WordBeguiler 算法伪代码如下。

输入:原始文本 x,真实标签 y,分类器 $F(\cdot)$,修改策略 $T(\cdot)$,扰动最大阈值 σ
输出:对抗样本 x^*
inititialize:$x^* \leftarrow x$
for s_i in document x do
　$Cs_i = F_y(s_i)$;
end for
Sordered←Sort(sentences) according to Cs_i;
Delete sentences in Sordered if $F(s_i) \neq y$;
for s_i in Sordered do
　for w_j in s_i do
　　Compute I_{w_j} according to Eq.8;
　end for
　Wordered←Sort(words) according to I_{w_j};
　for w_j in Wordered do
　　$w_j = T(w_j)$;
　　if $F(x^*) \neq F(x)$ and $\text{Cost}(x^*, x) \leqslant \sigma$
　　　return x^*;
　　end if
　end for
end for
return none

3　实验与结果

3.1　实验设置

目标模型:CNN[19] 由 300 维的嵌入层、三个卷积层和一个全连接层组成,卷积层由 256 个大小为 2,3,4 的卷积核组成,步长为 1;BiLSTM[20] 由一个 300 维的嵌入层,一个双向 LSTM 层和一个全连接层组成,双向 LSTM 层的前向和后向分别由 64 个 LSTM 单元组成。

数据集:选用四个中文数据集进行评估,如表 4 所示。

表 4　实验数据集

数据集	任务类型	类别数	训练集	测试集	文本长度
携程酒店评论	情感分析	2	12 000	3 000	135
京东购物评论	情感分析	2	35 000	5 000	32
垃圾信息	信息分类	2	90 000	10 000	56
新闻标题	新闻分类	10	90 000	10 000	23

3.2　实验方法比较

为了验证 WordBeguiler 算法的有效性,在情感分析数据集上分别与 DeepWordBug 和 WordHanding 两种攻击算法进行比较,实验中经过调试设置参数 $K=30$,修改策略中概率权重依次为字词替换 0.6、字符插入 0.2 和次序修改 0.2。携程酒店、京东购物、垃圾信息数据集允许最大修改阈值为 30,新闻分类最大修改阈值为 11。实验结果如表 5 和表 6 所示。

表 5　在情感分析任务上验证算法 WordBeguiler 的有效性

(a)CNN 模型

数据集	攻击方法	原始样本准确率/%	对抗样本准确率/%	降幅/%
携程	DeepWordBug	91.27	75.46	15.81
酒店	WordHanding		67.24	24.03
评论	WordBeguiler		38.22	53.05
京东	DeepWordBug	90.85	72.62	18.23
购物	WordHanding		68.16	22.69
评论	WordBeguiler		45.32	45.53

(b)BiLSTM 模型

数据集	攻击方法	原始样本准确率/%	对抗样本准确率/%	降幅/%
携程	DeepWordBug	91.26	68.52	22.74
酒店	WordHanding		60.77	30.49
评论	WordBeguiler		40.76	50.50
京东	DeepWordBug	93.04	70.28	22.76
购物	WordHanding		63.81	29.23
评论	WordBeguiler		48.43	44.61

表 6　在其他任务上评估 WordBeguiler 性能

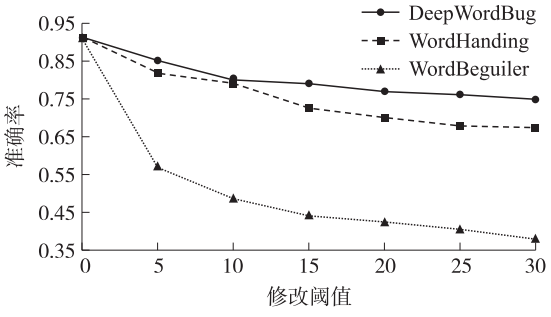
数据集	目标模型	原始准确率/%	WordBeguiler 准确率/%	降幅/%
垃圾信息	CNN	99.91	48.04	51.87
数据集	BiLSTM	99.84	47.91	51.93
新闻标题	CNN	91.52	52.35	39.17
数据集	BiLSTM	91.23	51.75	39.48

在四个数据集上分析了对 CNN 和 BiLSTM 模型的攻击效果。从表 5 可以看出,在情感分析数据集上与基线方法 DeepWordBug 和 WordHanding 相比,本文提出的 WordBeguiler 算法最高可使准确率降幅达 53.05%,明显优于基线方法。从表 6 可以看出,在垃圾信息和新闻标题数据集上 WordBeguiler 算法平均可使模型降低 45%左右的准确率,表明了其在多场景任务中的有效性与通用性。

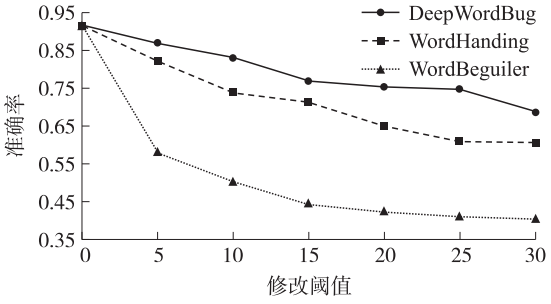
为了验证对抗样本检测的准确率与修改阈值 m 之间的关系,从两个情感分类数据集中分别选取了长度大于 120 个字的 1000 条样本,通过调整不同的修改幅度生成对应的对抗样本。

在携程酒店数据集和京东购物评论数据集上得出 CNN 和 BiLSTM 的检测准确率随修改阈值 m 的变化曲线,如图 3 和图 4 所示。

随着修改阈值 m 的增加,模型的检测准确率逐渐降低,意味着对输入序列中的个别关键字进行修

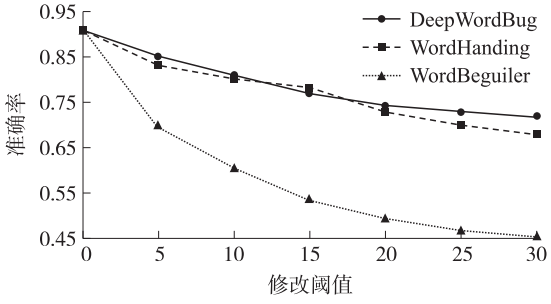


(a) CNN模型对比实验

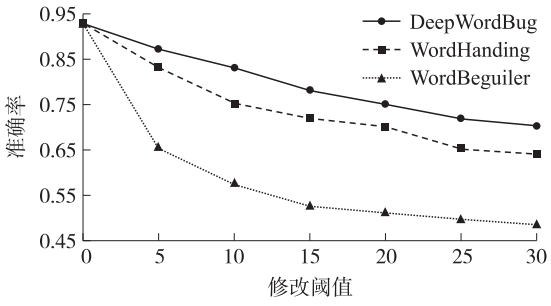


(b) BiLSTM模型对比实验

图 3 携程酒店评论数据集不同阈值实验结果



(a) CNN模型对比实验



(b) BiLSTM模型对比实验

图 4 京东购物评论数据集不同阈值实验结果

改,WordBeguiler 就可以生成对抗样本。与基线相比,当修改阈值达到 15 个字左右时,攻击效果趋于稳定状态,表明 WordBeguiler 算法大大地降低了扰动率,提高了文本的可读性。

3.3 对抗样本质量度量

图像领域中度量样本质量的方法典型的有 L_0, L_2 和 L_∞ 范数,但是由于图像与文本存在不同,因此上述度量方式不能直接应用在文本中。故使用词移距离(Word Mover's Distance,WMD)度量两个文本文档之间的距离,用于判断两个文本之间的相似度,即 WMD 距离越大文本相似度越小,WMD 距离越小相似度越大。从生成的对抗样本中随机选取 2 000 个样本进行实验,实验结果如图 5 所示。

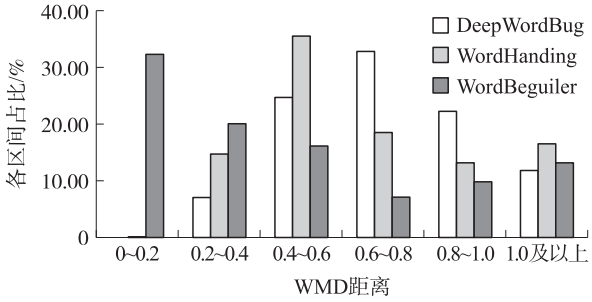


图 5 不同 WMD 距离区间内样本数量占总样本的比例

与基线方法相比,WMD 距离小于 0.2 的对抗样本占总样本的 32.6%,在 0~0.6 区间的对抗样本总占比为 69.28%,高于基线方法,说明由 WordBeguiler 算法生成的对抗样本与原样本有较高的相似度。表 7 是原始样本与 WordBeguiler 算法生成的对抗样本的示例,很好地保留了语义且可读性好。

表 7 原始样本与对抗样本示例

原始样本:服务非常好,环境也很舒适,只是酒店的餐饮方面一般般。	积极
对抗样本:服务匪常\耗,环境也很舍予 shi,只是酒店的餐饮方面一般般。	消极
原始样本:特价的时候买的,一直在用,还是很不错的。	积极
对抗样本:价特的时候埋的,一直在用,还是恨不! 挫的。	消极

3.4 可转移性评估

为了验证由 WordBeguiler 算法生成的对抗样本具有可转移性,分别使用 BiLSTM 和 CNN 模型生成对抗样本去攻击其他模型(包括 BERT[22] 模型)的实验,如表 8 和表 9 所示。实验结果表明,WordBeguiler 算法生成的对抗样本能够有效地转移到其他模型上,且准确率降幅达 30%左右。

表 8 使用 BiLSTM 模型生成对抗样本攻击其他模型

数据集	模型	原始准确率/%	WordBeguiler	
			准确率/%	降低幅度/%
携程酒	CNN	91.27	55.43	35.84
店评论	BERT	90.13	56.76	33.37
京东购	CNN	90.85	59.21	31.64
物评论	BERT	91.24	60.68	30.56

表 9　使用 CNN 模型生成的对抗样本攻击其他模型

数据集	模型	原始准确率/%	WordBeguiler	
			准确率/%	降低幅度/%
携程酒	BiLSTM	91.26	56.70	34.56
店评论	BERT	90.13	57.88	32.25
京东购	BiLSTM	93.04	61.19	31.85
物评论	BERT	91.24	62.88	28.36

4　结语

本文提出了在黑盒条件下面向中文文本分类的对抗样本生成方法 WordBeguiler,诱使模型做出错误决策。该方法首先结合汉字字形、字音特征为每个汉字构建对抗搜索空间,以提高对抗样本质量,设计了新的扰动定位方式寻找影响分类的重要字或词组,并根据概率权重选取的方法确定修改策略生成对抗样本。WordBeguiler 算法生成的对抗样本可以使 CNN 和 BiLSTM 模型的准确率降幅最高达 53.05% 和 50.50%,转移到其他模型上准确率降幅达 30%左右,攻击效果均优于其他攻击方法。此外,字词重要性计算函数与修改策略还能进一步优化与改进,今后的工作重心会集中到以上方面,并对如何增强文本分类模型的鲁棒性问题展开更深入的探索与研究。

参考文献:

[1] Lu D,Weng Q. A Survey of Image Classification Methods and Techniques for Improving Classification Performance [J]. International Journal of Remote Sensing,2007,28(5):823-870.

[2] Deng L, Hinton G, Kingsbury B. New Types of Deep Neural Network Learning for Speech Recognition and Related Applications: An Overview[C]//2013 IEEE International Conference on Acoustics,Speech and Signal Processing,Vancouver,BC,Canada,IEEE,2013:8599-8603.

[3] Young T, Hazarika D, Poria S, et al. Recent Trends in Deep Learning Based Natural Language Processing[J]. IEEE Computational Intelligence Magazine,2018,13(3):55-75.

[4] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing Properties of Neural Networks [C]//2nd International Conference on Learning Representations,Banff,AB,Canada,2014:arXiv:1312.6199.

[5] Goodfellow I J,Shlens J,Szegedy C. Explaining and Harnessing Adversarial Examples[EB/OL]. (2015-03-20). https://arXiv.org/abs/1412.6572.

[6] Madry A, Makelov A, Schmidt L, et al. Towards Deep Learning Models Resistant to Adversarial Attacks[C]//2nd International Conference on Learning Representations,Banff,AB,Canada,2014:arXiv:1312.6199

[7] Wiyatno R,Xu A Q. Maximal Jacobian-Based Saliency Map Attack[EB/OL]. (2018-08-23). https://arXiv.org/abs/1808.07945.

[8] Ng H W,Nguyen V D,Vonikakis V,et al. Deep Learning for Emotion Recognition on Small Datasets Using Transfer Learning[C]//2015 ACM International Conference on Multimodal Interaction,Seattle,WA,USA,2015:443-449.

[9] Johnson R,Zhang T. Effective Use of Word Order for Text Categorization with Convolutional Neural Networks [C]//2015 Conference of the North America Charter of the Association for Computational Linguistics:Human Language Technologies,Denver,CO,USA,2015:103-112.

[10] Singh S P,Kumar A,Darbari H,et al. Machine Translation Using Deep Learning:An Overview[C]//2017 International Conference on Computer, Communications and Electronics (Comptelix),Jaipar,India,IEEE,2017:162-167.

[11] Sharma Y,Gupta S. Deep Learning Approaches for Question Answering System[J]. Procedia Computer Science,2018,132:785-794.

[12] 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,MD,USA,2016:49-54.

[13] Li L,Ma R,Guo Q,et al. Bert-Attack:Adversarial Attack Against Bert Using Bert [C]//2020 Conference on Empirical Methods in Nature Language Processing,Online,2020:6193-6202.

[14] 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]//34th AAAI Conference on Artificial Intelligence,New York,NY,USA,2020:8018-8025.

[15] 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, CA, USA,2018:50-56.

[16] Garg S,Ramakrishnan G. BAE:BERT-Based Adversarial Examples for Text Classification[C]//2020 Conference on Empirical Methods in Nature Language Processing,Online,2020:6174-6181.

[17] 王文琦,汪润,王丽娜,等. 面向中文文本倾向性分类的对抗样本生成方法[J]. 软件学报,2019,30(8):2415-2427.

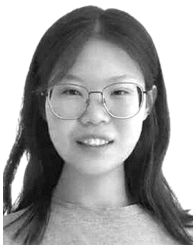
[18] 仝鑫,王罗娜,王润正,等. 面向中文文本分类的词级对抗样本生成方法[J]. 信息网络安全,2020,20(9):12-16.

[19] Kim Y. Convolutional Neural Networks for Sentence Classification [C]//The 2014 Conference on Empirical Methods in Natural Language Processing, Doha, Qatar, 2014:1746-1751.

[20] Graves A, Schmidhuber J. Framewise Phoneme Classification with Bidirectional LSTM and Other Neural Network Architectures[J]. Neural Networks, 2005, 18(5-6): 602-610.

[21] Hochreiter S, Schmidhuber J. Long Short-Term Memory [J]. Neural Computation, 1997, 9(8): 1735-1780.

[22] Devlin J, Chang M W, Lee K, et al. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding[C]//2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2019:4171-4186.



弓 燕(1997—),女,汉族,内蒙古呼和浩特人,内蒙古科技大学信息工程学院硕士研究生,主要研究方向为信息安全及自然语言处理,3174917341@qq.com;

张晓琳(1966—),女,汉族,内蒙古包头人,博士,教授,硕士生导师,主要研究方向为机器学习及信息安全,zhangxl@imust.cn。