软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2019,30(8):2415−2427 [doi: 10.13328/j.cnki.jos.005765]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

面向中文文本倾向性分类的对抗样本生成方法*

王文琦[1,2], 汪 润[1,2], 王丽娜[1,2], 唐奔宵[1,2]

[1](空天信息安全与可信计算教育部重点实验室(武汉大学),湖北 武汉 430072)

[2](武汉大学 国家网络安全学院,湖北 武汉 430072)

通讯作者: 王丽娜, E-mail: lnawang@163.com

摘 要: 研究表明,在深度神经网络(DNN)的输入中添加小的扰动信息,能够使得 DNN 出现误判,这种攻击被称为对抗样本攻击.而对抗样本攻击也存在于基于 DNN 的中文文本的情感倾向性检测中,因此提出了一种面向中文文本的对抗样本生成方法 WordHanding.该方法设计了新的词语重要性计算算法,并用同音词替换以生成对抗样本,用于在黑盒情况下实施对抗样本攻击.采用真实的数据集(京东购物评论和携程酒店评论),在长短记忆网络(LSTM)和卷积神经网络(CNN)这两种 DNN 模型上验证该方法的有效性.实验结果表明,生成的对抗样本能够很好地误导中文文本的倾向性检测系统.

关键词: 中文文本;对抗样本;深度学习模型;评分函数;黑盒

中图法分类号: TP309

中文引用格式: 王文琦,汪润,王丽娜,唐奔宵.面向中文文本倾向性分类的对抗样本生成方法.软件学报,2019,30(8):2415−2427. http://www.jos.org.cn/1000-9825/5765.htm

英文引用格式: Wang WQ, Wang R, Wang LN, Tang BX. Adversarial examples generation approach for tendency classification on Chinese texts. Ruan Jian Xue Bao/Journal of Software, 2019,30(8):2415−2427 (in Chinese). http://www.jos.org.cn/1000-9825/5765.htm

Adversarial Examples Generation Approach for Tendency Classification on Chinese Texts

WANG Wen-Qi[1,2], WANG Run[1,2], WANG Li-Na[1,2], Tang Ben-Xiao[1,2]

[1](Key Laboratory of Aerospace Information Security and Trusted Computing (Wuhan University), Ministry of Education, Wuhan 430072, China)

[2](School of Cyber Science and Engineering, Wuhan University, Wuhan 430072, China)

Abstract: Studies have shown that the adversarial example attack is that small perturbations are added on the input to make deep neural network (DNN) misbehave. Meanwhile, these attacks also exist in Chinese text sentiment orientation classification based on DNN and a method “WordHandling” is proposed to generate this kind of adversarial examples. This method designs a new algorithm aiming at calculating important words. Then the words are replaced with homonym to generate adversarial examples, which are used to conduct an adversarial example attack in black-box scenario. This study also verifies the effectiveness of the proposed method with real data set, i.e. Jingdong shopping and Ctrip hotel review, on long short-term memory network (LSTM) and convolutional neural network (CNN). The experimental results show that the adversarial examples in this study can mislead Chinese text orientation detection system well.

Key words: Chinese text; adversarial examples; deep learning models; score function; black box

* 基金项目: 国家自然科学基金(61876134); 国家重点研发计划(2016YFB0801100); 中央高校基本科研业务费专项资金(2042018kf1028)

Foundation item: National Natural Science Foundation of China (61876134); National Key Research and Development Program of China (2016YFB0801100); Fundamental Research Funds for the Central Universities (2042018kf1028)

本文由“面向自主安全可控的可信计算”专题特约编辑贾春福教授推荐.

收稿时间: 2018-05-31; 修改时间: 2018-09-21; 采用时间: 2018-12-13; jos 在线出版时间: 2019-03-28

CNKI 网络优先出版: 2019-03-29 09:47:18, http://kns.cnki.net/kcms/detail/11.2560.TP.20190329.0947.016.html

基于深度神经网络(deep neural network,简称 DNN)的机器学习方法已被广泛地应用于许多领域,如计算机视觉[1,2]、语音识别[3]、自然语言处理[4−9]、恶意软件检测[10−12]等,但 DNN 在上述应用中都面临着对抗样本攻击的威胁.对抗样本是指在正常的样本中通过有目地性添加少量的扰动信息,使得基于 DNN 模型的系统出现误判[13,14].Szegedy 等人[14]已证实了对于一些机器学习模型,包括在多方面表现很好的神经网络模型,在遭受对抗样本的攻击时都表现出明显的脆弱性.

对抗样本最初发现在基于 DNN 的图像识别中,如自动驾驶中,攻击者对路标图像进行修改使得车辆识别系统把左转判别为右转,存在极大的安全隐患.而对抗样本不仅出现在图像领域,本文发现,在基于 DNN 的中文文本倾向性检测中也存在对抗样本攻击的问题.如判断网络中传播的文本信息是正常还是异常时,攻击者可以利用对抗样本生成的方法对异常信息进行处理,使处理之后的异常信息被检测系统误判为正常信息,“欺骗”系统的检测,使得异常信息扩散.或者是把大量恶意评论“伪装”成正常评论散播,影响人们对人和事的情感倾向或对物的购买欲望.如图 1 所示,某商品经过推荐系统宣传销量增加,然而受商业竞争对手雇佣的攻击者把针对该商品的恶意信息进行修改生成对抗样本,使基于 DNN 的恶意信息检测系统产生错误的判别而未拦截.这导致人们被大量负面信息影响,对该商品由购买倾向变为不买,最终使得产品销量降低.这些安全性问题自然引起了人们对深度学习模型鲁棒性的关注,而对深度学习模型具有威胁性的对抗样本生成过程的研究同样有着重要意义,其有助于分析基于深度学习模型系统存在的安全问题,有助于建立针对此类攻击的检测防御工作[15].

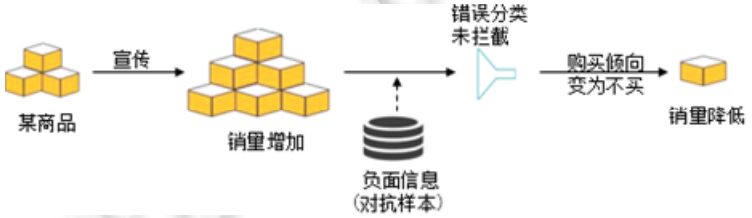


Fig.1 Impacts of adversarial example attack

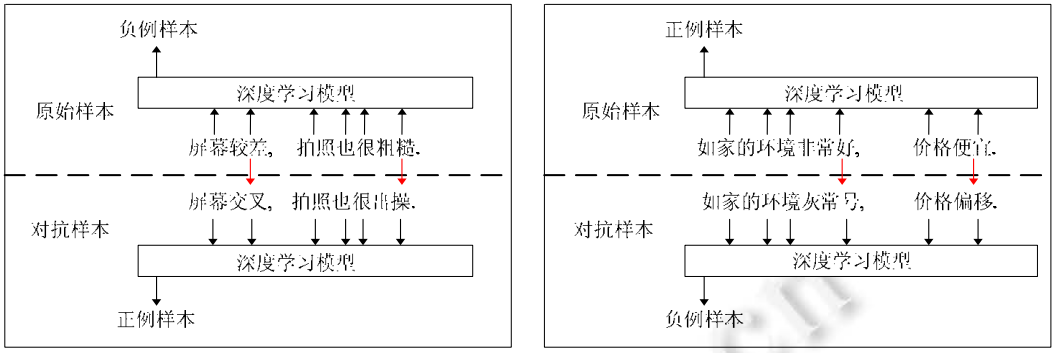
图 1 对抗样本攻击带来的影响

按攻击者对目标模型的了解程度,可以将对抗样本攻击分为白盒、黑盒可探测以及黑盒不可探测攻击:白盒情景下,攻击者对模型完全掌握,包括模型的类型、结构、所有参数及权重值;黑盒可探测情景下,攻击者对目标模型部分了解,但能对模型进行探测或查询,如通过输入观察输出结果;黑盒不可探测情景下,攻击者对目标模型了解有限或完全没有,构建对抗样本时不允许对目标模型进行探测查询.在现实中,黑盒攻击比白盒更为实际,但如今,大多数文本类型对抗样本的生成和攻击是以白盒为前提假设[16−19],黑盒条件下的研究很少[20,21].此外,目前的研究都是针对英文数据,一般的修改方法包含对输入中的字母进行操作,如插入、删除、相邻字母位置互换等.该方法不适用于中文数据,因为每个汉字是一个独立单元,不可拆分;而使用邻近词[19]、添加标点的方式,这些方法会改变原输入语句的意思,也不适用.因此,如何对中文文本数据处理生成对抗样本,在更实际的黑盒条件下实现针对长短记忆网络(long short-term memory,简称 LSTM)和卷积神经网络(convolutional neural network,简称 CNN)这两种常见的情感分析模型的对抗样本攻击,是本文要解决的问题.

本文提出了一种面向中文文本的对抗样本生成方法 WordHandling,该方法不需要直接目标网络的参数信息,其通过设计一个新的算法计算文本中影响分类的重要词语,用同音词或词组替换的方法修改原始数据生成对抗样本,有效地实现了针对深度神经网络模型的黑盒攻击.图 2 所示为利用 WordHandling 生成中文对抗样本的例子.

图 2(a)和图 2(b)虚线上半部分是原始样本,下半部分是生成的对抗样本.从正常的情感倾向看,图 2(a)和图 2(b)中的原始样本分别为负面评论与正面评论.可以看出,原始的输入样本能够被深度神经网络模型正确的分类.但仅仅对输入的中文文本进行些许改动,就能干扰深度学习倾向检测系统,使其产生错误的倾向判断,把正面评论判断为负面评论,或者是负面评论判别为正面评论.同时,由于句子修改前后的含义内容变化微小,人仍

然能够通过句子的语义或语音来理解修改后的语句.



(a) 修改负面评论例子 (b) 修改正面评论例子

Fig.2 Examples of generated adversarial example es by WordHandling

图 2 WordHandling 生成的对抗样本样例

本文的主要贡献有:

(1) 提出了一种中文对抗样本生成方法,只需要对输入的中文文本进行些许修改,且不需要知道目标模型的参数,即能生成对抗样本,可用于干扰文本的情感倾向分类.

(2) 本文设计了一种新的词语重要性计算方法,利用该方法,以较小的代价对中文文本数据进行修改,有效改变 DNN 模型对修改后样本的倾向分类.

(3) 提出的方法在真实的数据集上进行实验,使用 LSTM 和 CNN 模型对生成的对抗样本做倾向判别,倾向判别准确率平均下降 29%和 22%,而对输入的中文文本平均的修改幅度为 14.1%,实验结果证明了文中所提的 WordHandling 对抗样本生成方法的有效性.

本文第 1 节是相关工作的介绍.第 2 节是背景知识.第 3 节对本文提出的对抗样本生成算法进行详细描述.第 4 节为实验设置以及结果的分析讨论.第 5 节为本文的总结.

1 相关工作

传统机器学习的安全性方面已进行了许多研究工作,如针对机器学习模型形成的不同类型攻击方法的总结及相应的防御措施[22].其中提出的一些攻击方法,包括“污染”攻击[23,24]和回避攻击[25−27],威胁着已部署的基于机器学习模型的系统.而在各方面表现很好的深度神经网络模型也同样存在着安全问题,如精巧制作对抗样本,能够误导基于 DNN 的分类系统,这种新的攻击方式及其针对性的防御引起越来越多的关注.就对抗样本攻击而言,现已有多种生成对抗样本的方法用于对抗 DNN,如 FGSM(fast gradient sign method)[28−30],JSMA(jacobian-based saliency map attack)[31],Deepfool[32]等.但上述方法多是针对图像类型,并不能直接应用于文本领域.原因在于图像是连续的,文本是离散的且有词序限制;另一方面,文本和图像的距离度量标准并不一致.

尽管不同于图像领域,文本方面的对抗样本研究工作也已有相应的进展,Liang 等人[17]通过对输入数据的词向量梯度计算决定向文本中插入、删除、修改的内容,在哪里插入以及如何进行修改,但是该方法需要知道模型参数,不适合本文的黑盒场景,而且插入整个语句的方法可能会改变输入内容的意思.Ebrahimi 等人[18]使用同义词来替代原词,其中加入了严格的限制条件.但是该方法在仅改动一两个词就可生成对抗样本的情况下,得到的对抗样本数量十分稀少.Papernot 等人[19]使用 LSTM,把随机选取的词在嵌入(embedding)层的词向量用向量空间中最邻近的词向量替代,映射到输入中就可能会生成与原词完全不相关的词来代替,不能保证语义上的相似.此外,Gao 等人[21]在黑盒条件下生成文本类型的对抗样本,他们提出了 DeepWordBug 算法,根据黑盒条件下观察模型的输出结果,设计词语重要性计算函数,找出文本中的关键单词,并对单词的字母进行插入、删除、

取代、前后字母交换位置等方式修改以生成对抗样本,但修改的方法不适用于中文数据.上述方法使用的实验数据是英文文本,并没有使用中文实验数据的研究工作.本文在第 3 节介绍针对中文文本数据生成对抗样本的算法.

2 背景知识

2.1 情感倾向性分类

情感倾向性分析的目的是利用机器提取人们在文本中对某事物或某人表现的态度,从而发现潜在的问题来改进和预测.主要的分析内容是对特定的人或事物带有主观色彩的偏好和倾向,如喜欢、讨厌、好、坏等.而分类任务是把自然语言编写的文档通过深度学习模型自动地划分到预先定义的类别中[33].在文本领域,确定产品或评论中的情感倾向的情感分类也是一种文本分类任务.文本分类典型的方法是用词袋向量(bag-of-word vectors)表示文本数据,使用 SVM(support vector machine)等传统机器学习模型并不能保证词在文本中的词序,而词序的缺失对文本的情感分类有很大影响.Johnson 等人[33,34]分别使用 CNN 和 LSTM 在保证词序的前提下对词向量处理,进行情感分类.而在实际生活中,情感倾向性分类可用于评论筛选、信息过滤等工作.

2.2 卷积神经网络CNN

卷积神经网络 CNN[35]是一种前馈神经网络,其包含卷积层与池化层.卷积操作的作用是突出特征,将更明显的特征提取出来,在 CNN 中,卷积的操作可以用公式(1)来表示.

$$c_i = f(\sum w \cdot x + bias) \tag{1}$$

该公式表示卷积核与卷积区域的点乘和,然后同偏置求和后激活.池化也称为子采样,其可看作为一种特殊的卷积过程,常有均值和最大值子采样两种形式.卷积和池化能简化模型的复杂度,减少其中的参数.

CNN 在多个领域取得了成功,在图像处理方面,其特有的卷积、池化结构能够提取图像中各种不同程度的纹理、结构,并最终结合全连接网络实现信息的汇总和输出.而在短文本分析任务中,句子的长度有限、结构紧凑、能够独立表达意思的特点,使得 CNN 在处理这一类问题上成为可能[36],因而其也能用在文本倾向性检测中.

2.3 循环神经网络RNN

面对时序相关的输入数据时,传统的神经网络表现不佳,其直进直出的特性决定这只能使用当前输入而不能利用之前的数据信息.循环神经网络(recurrent neural network,简称 RNN)很好地解决了这一问题,RNN 是一个循环网络,能够很好地存储信息,其网络结构如图 3 所示.图 3 对 h 处的循环展开,可看出,单元 h_t 不仅受当前输入 x_t 的影响,还受到 h_t 之前单元的影响.$W1$~$W3$ 表示权值.

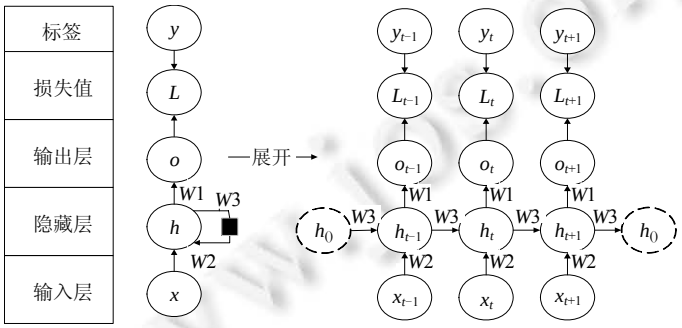


Fig.3 Network structure of RNN

图 3 RNN 网络结构

但当前输入信息与之前的输入关联度有长有短,随着信息关联度的变长,RNN 不能学习这些信息之间的联系而失去效用.为解决这个长期依赖问题,LSTM 因此被提出.长短记忆网络 LSTM[37]是一种特殊的循环神经网

络 RNN,由 Sepp Hochreiter 和 Jürgen Schmidhuber 在 1997 年提出,并加以完善与普及,其能够学习到长期依赖关系,可以对之前的输入有选择的记忆,从而有助于判断当前输入.LSTM 在各类任务上表现良好,包括情感倾向性检测,因此被广泛使用[38−40].

3 WordHandling 算法

3.1 前提设定

本文预先训练一个 LSTM 替代模型,把一个文本数据作为输入,经过 LSTM 模型后,会输出一个分数 s,根据预先设定的阈值与 s 的比较来判别该输入的类别倾向.一般黑盒条件下,观察模型的结果仅能得到输入数据所属的类别标签 y,而本文通过 LSTM 替代模型,可获取且仅需获取输出的判别分数 s.

① 设定训练集中的正负样本评论数据分别标记为 1 和 0,倾向分类判别阈值用λ表示.当$s>\lambda$时,该输入被判别为正样本(positive);当$s\leqslant\lambda$时,被判别为负样本(negative).

② 原始样本分词后,依次输入得到各自的分数,认为 s 在β到α之间的词语为偏中性,不带情感倾向或倾向微弱;$s>\alpha$,偏正面;$s<\beta$,偏负面.

③ 名词不包含明显的情感倾向,可排除在修改序列外.

④ 对于汉字中不存在同音字的词,如“嗯”“命”等,但这些字并不影响 WordHandling.原因在于,不存在同音字的汉字稀少,并不存在明显的情感倾向且可以用谐音词代替.这些词计算得到的重要性程度弱于其他情感倾向较大的词,如正面倾向的“喜欢”“好”,负面倾向的“坏”“差”“无用”等.

3.2 总体描述

自然语言处理的相关应用对输入文本中的某些词具有高度的敏感性,对敏感词进行修改在较大程度上能够改变分类器对输入样本的类别倾向判断[5].而敏感词也即是语句中的重要词(关键词或贡献度大的词),其重要性或贡献度则是对输入数据类别倾向判断影响程度的一个度量,重要词被修改之后很大程度上会改变原始输入的情感类别倾向.因此,本文对语句中情感倾向性词语进行修改以生成对抗样本的流程,如图 4 所示.

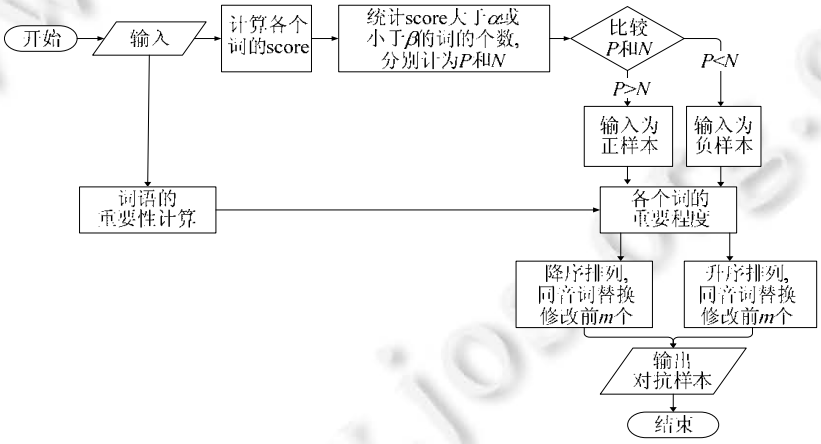


Fig.4 Adversarial examples generation process in this paper

图 4 本文对抗样本生成的流程

其过程描述如下:

① 对输入文本进行分词处理,依次输入计算各个词或词组的分数 score.

② 统计 score 大于α(偏正面)和小于β(偏负面)的词的个数,分别记为 P 和 N.

③ 比较 P 和 N 的大小:若 $P>N$,则认为输入的数据偏正面;反之,则认为输入的数据偏负面.

④ 利用词语重要性计算函数计算输入文本数据中各个词或词组的重要程度.

⑤ 若输入数据偏正面,对各个词的 score 按降序排列;若输入数据偏负面,则按升序排列.

⑥ 对排序在前 m 的词或词组(剔除名词后的)用同音词替换修改得到对抗样本,m 为修改文本中词或词语的次数或幅度.

3.3 输入文本中词或词组重要程度的计算

为保证修改后得到的文本的可读性和有效性,对文本中重要词或词组进行修改并控制改动的幅度是必须的.但哪些是重要词语以及如何确定,是本节将要解决的问题.对重要词语的计算则需要使用词语重要性计算函数,本文设计了新的方法来进行计算.

- Delete Score (DS)

对输入样本 $\boldsymbol{X}$ 进行分词得到 $\boldsymbol{X}=[x_1,x_2,x_3,\ldots,x_n]$,$n$ 表示分词的长度,对序列中的第 i 个词语,计算整句输入和移除第 i 个词语后的输入分数的差值:

$$DS(x_i)=F(x_1,\ldots,x_{i-1},x_i,x_{i+1},\ldots,x_n)-F(x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_n) \tag{2}$$

- Forward Score (FS)

该函数计算的是序列中第 i 个词语的预测分数,通过计算输入中前 i 个词语和前 i–1 个词语分数的差值:

$$FS(x_i)=F(x_1,x_2,\ldots,x_{i-1},x_i)-F(x_1,x_2,\ldots,x_{i-1}) \tag{3}$$

其中,假定当 i=1 时,$FS(x_i)$=0.

- TF-IDF Score

TF-IDF 提取输入数据中的关键词,提取的关键词中可能含有带情感倾向的词:

$$TF=\text{词语在当前输入样本数据中出现的次数/当前输入样本数据的总词数} \tag{4}$$

$$IDF=\log(\text{输入样本数据的总数目/(包含该词的输入样本数据的数目+1)}) \tag{5}$$

$$TF\text{–}IDF=TF\cdot IDF \tag{6}$$

移除计算得到的关键词中的名词词语,找到这些关键词在 $\boldsymbol{X}=[x_1,x_2,x_3,\ldots,x_n]$中的位置索引,把长度为 n 的零向量 M 中相应位置的数值 0 用 $TF\text{-}IDF$ 值替代,得到向量 $\boldsymbol{M}=[m_1,m_2,m_3,\ldots,m_n]$.则词语的 TIS score 为

$$TIS(x_i)=\frac{m_i-m_{\min}}{m_{\max}-m_{\min}} \tag{7}$$

其中,x 为向量中数值,$m_{\max}$ 和 $m_{\min}$ 分别为 M 中所有数据最大值和最小值.最终,通过词语重要性计算函数对输入样本 $\boldsymbol{X}$ 中的第 i 个词语重要程度进行计算:

$$score(x_i)=DS(x_i)\cdot w_1(i)+FS(x_i)\cdot w_2+TIS(x_i)\cdot w_3(i) \tag{8}$$

$$w_1(i)=1/(1+\mathrm{e}^{-F(x_i)}) \tag{9}$$

$$w_3(i)=\begin{cases}1, & F(x_i)>\lambda\\ -1, & F(x_i)\leqslant\lambda\end{cases} \tag{10}$$

其中,w_2 是一个超参数.

3.4 输入文本中重要词或词组的修改

近年的研究工作中,文本类型对抗样本生成使用的是英文数据,采用的方法是对输入中的英文词语进行直接修改,方法包括词语的替换、插入、删除及单词中相邻字母位置交换等.但这些方法对于中文文本数据而言并不适用,对中文文本中的词语进行插入、删除,极大多数情况会改变原语句的意思,而单个中文汉字则并不存在相邻字母位置交换的情况.因而在输入层面而非词向量嵌入层和语义层上对中文样本进行直接修改,重要词语替换的方式具有可行性.但简单的随机替换也同样存在改变原数据的意思的问题,而本文则提出一个有效且可行的方法,使用词语重要性计算函数计算语句中各个词的重要程度,对重要性高的词语通过获取它们的同音词或词组来代替原词语,生成对抗样本来迷惑深度学习模型,提高其误检率.这样做的好处是:人仍然能够通过句子的上下文或词的谐音来理解句子的含义,较大程度上保留原样本数据的内容,同时又能够“避开”学习模型

的检测.

算法. WordHandling 算法.

输入:$\boldsymbol{X}=x_1x_2x_3\ldots x_n$,修改幅度 m,词语重要性计算函数 S,排序函数 $Rank$,转换函数 T,功能函数 F.

输出:$\boldsymbol{X}'$.

1. $\boldsymbol{X}=[x_1,x_2,x_3,\ldots,x_n]$
2. **for** x in $\boldsymbol{X}$:
3. $\quad \boldsymbol{Score}_i=F(x)$
4. $\quad \boldsymbol{score}_i=S(x)$
5. $label=\boldsymbol{Judge}(Score)$
6. **for** s in $score$:
7. $\quad$ **if** $label$==‘pos’:
8. $\quad\quad index=\boldsymbol{Rank\_descending}(s)$
9. $\quad$ **elif** $label$==‘neg’:
10. $\quad\quad index=\boldsymbol{Rank\_ascending}(s)$
11. **for** i in $range(m)$:
12. $\quad x'_{index\ i}=\boldsymbol{T}(x_{index\ i})$
13. **return** $\boldsymbol{X}'$

算法说明:对于输入的中文数据进行分词,$\boldsymbol{X}$ 中的元素可能是单个汉字,也可能是词组;然后计算 $\boldsymbol{X}$ 中每个元素的重要性,判断输入样本情感偏向性;按情感偏向的正或负来对词语的重要性进行降或升序排列,选取重要性排名前 m 个重要词进行修改得到生成的对抗样本.以酒店评论为例,一般只包含正面评论和负面评论两类,输入样本经过算法步骤 5 初步判断情感倾向后,若判定为正面评论,对各个词的重要程度降序排列,偏向正面的词或词组的分数排在前列,对这些词进行修改能较大程度上改变样本的情感倾向,使模型对修改后的样本类别产生不同判定,但与原样本的内容含义差异微小.

在本文中,修改幅度 m 是动态变化的.在对抗样本生成的过程中,依次对排序后的词进行修改,判断修改前后样本的类别是否改变,直到影响原输入样本情感倾向的词(即 $score>\alpha$ 或 $score<\beta$)的词被全部修改为止.对于定长的输入样本而言,修改的幅度 m 越大,其改动后类别的变化可能性越大.第 4.2 节中给出深度学习模型对给定长度的数据以不同修改幅度生成对抗样本的检测准确率的变化情况.

4 实验设置与结果分析

4.1 实验设置

本文在不同的深度神经网络模型上进行对抗样本的有效性验证,采用的数据集形式见表 1.

Table 1 Experimental data set

表 1 实验数据集

项目	携程酒店评论数据	京东购物评论数据
任务类型	倾向性分类	倾向性分类
分类数目	2	2
训练集(条)	6 000	4 800
测试集(条)	3 000	3 000
文本平均长度(字)	108	37
文本中值长度(字)	146	26

实验所使用的是公开的携程酒店评论数据和京东购物评论数据.对于携程酒店评论数据集,从中选取正负面评价语料各 3 000 条作为训练样本,另随机取 3 000 条作为测试样本进行测试,选取的训练语料的平均长度为 108 字,中值长度为 146 字.对于京东购物评论数据集,从中选取 2 400 条五星评价的正面评论和 2 400 条一星评

价的负面评论作为训练样本,另随机选取 3 000 条评价为测试样本,选取的训练语料的平均长度为 37 字,中值长度为 26 字.实验中,训练和测试所使用到的数据其类标签是已知的,正面评论样本标记为 1,负面评论样本标记为 0,深度学习模型对样本的分类阈值λ数值为 0.5,α,β的数值分别为 0.6 和 0.4.

为了验证所提出的 WordHandling 的有效性,本文利用预先训练的 LSTM 替代模型对两个数据集中的测试集进行修改,生成相应的中文对抗样本,把这些对抗样本作为输入,对 LSTM,CNN 模型实施黑盒攻击.对于 WordHandling 实际效果的衡量则是通过深度学习模型对生成的对抗样本检测的准确率来体现,准确率越低,则攻击效果越好,生成的对抗样本更能有效地“避开”系统的检测,诱导模型产生错误的倾向分类.

4.2 实验方法的比较

本文在 LSTM 和 CNN 模型上验证所提出的 WordHandling 方法的有效性,关于模型检测准确率的结果见表 2(表中数值均为百分比),对比方法有 3 种:随机删除修改、随机选词替换以及 DeepWordBug[21].

Table 2 Verification of WordHandling validity on two datasets

表 2 两种数据集上验证 WordHandling 的有效性

(a) LSTM 模型

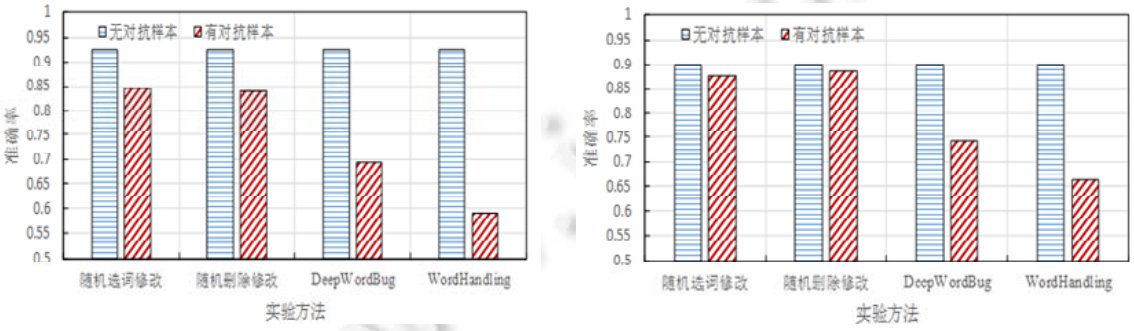
		对比方法						本文方法	
项目	无修改	随机选词修改		随机删除修改		DeepWordBug		WordHandling	
数据集	准确率	准确率	降低幅度	准确率	降低幅度	准确率	降低幅度	准确率	降低幅度
携程酒店评论	92.38	84.47	7.91	83.92	8.46	69.29	23.09	59.03	33.35
京东购物评论	88.51	83.15	5.36	80.62	7.89	67.53	20.98	61.25	27.26

(b) CNN 模型

		对比方法						本文方法	
项目	无修改	随机选词修改		随机删除修改		DeepWordBug		WordHandling	
数据集	准确率	准确率	降低幅度	准确率	降低幅度	准确率	降低幅度	准确率	降低幅度
携程酒店评论	90.12	87.51	2.61	88.43	1.69	74.16	15.96	66.42	23.70
京东购物评论	87.82	85.64	2.18	84.26	3.56	71.23	16.59	67.19	20.63

本文在不对输入文本进行较大规模改动及保证语义变化甚微的前提下,动态确定修改幅度 m.在本文的实验中,文本数据的修改幅度 m 平均占输入文本长度的 14.1%,近似等于文本长度的六分之一.对于携程酒店评论数据集,使用 WordHandling 算法生成对抗样本,并在深度学习模型上进行黑盒攻击,其中:LSTM 模型检测的准确率从 92.38%下降到 59.03%,CNN 模型检测的准确率从 90.12%下降到 66.42%;对于京东购物评论数据集,LSTM 模型检测的准确率从 88.51%下降到 61.25%,CNN 模型检测的准确率从 87.82%下降到 67.19%.

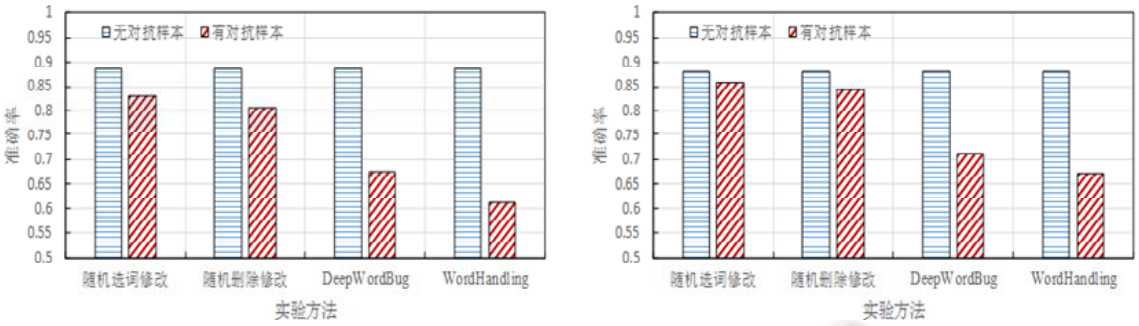
数据柱状图如图 5 和图 6 所示.



(a) LSTM 模型对比实验 (b) CNN 模型对比实验

Fig.5 Detection results of adversarial examples on Ctrip data in Table 2

图 5 表 2 中携程数据上对抗样本检测结果



(a) LSTM 模型对比实验 (b) CNN 模型对比实验

Fig.6 Detection results of adversarial examples on Jingdong data in Table 2

图 6 表 2 中京东数据上对抗样本检测结果

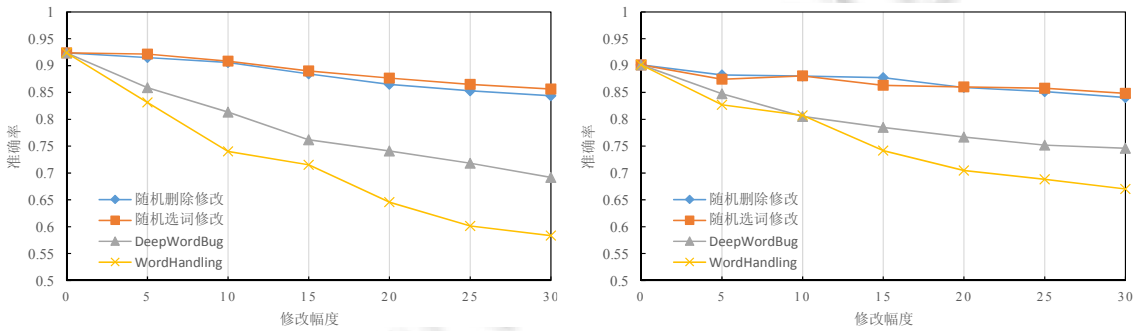
由表 2 可以看出,关键词语对输入数据类别倾向的影响较大,使用随机的方式对输入进行改动得到的结果并不理想;而且本文提出的 WordHandling 算法比 DeepWordBug 效果更佳.表 3 则是选取的若干原始样本和在其基础之上生成的对抗样本的例子,由表 3 可以看出,生成的对抗样本仍然能够通过语义上下文被人所理解,文本意思变化在可接受范围内.

Table 3 Examples of original examples and generated adversarial examples

表 3 原始样本和生成的对抗样本例子

原始样本:服务态度不好,换个房间都不给换,弄个最差的给住.	负面评价
对抗样本:服务态度部耗,换个房间都部给换,弄个醉岔的给住.	正面评价
原始样本:是非常不错的一家商务酒店,没有什么可以挑剔的了.	正面评价
对抗样本:是非常步挫的一家商务酒店,妹邮什么客衣跳题的了.	负面评价
原始样本:很一般,性价比很差.跟上海的快捷酒店相比,价格贵,服务差;窗户的高度太低,不小心会摔下去;令人匪夷所思的是,订的是大床房,但是床却是坏的;感觉很不好.	负面评价
对抗样本:很易班,性价比很岔.跟上海的快捷酒店相比,价格柜,服务岔;窗户的高度泰笛,不小心会摔下去;令人匪夷所思的是,订的是大床房,但是床却是踝的;感觉很部皓.	正面评价

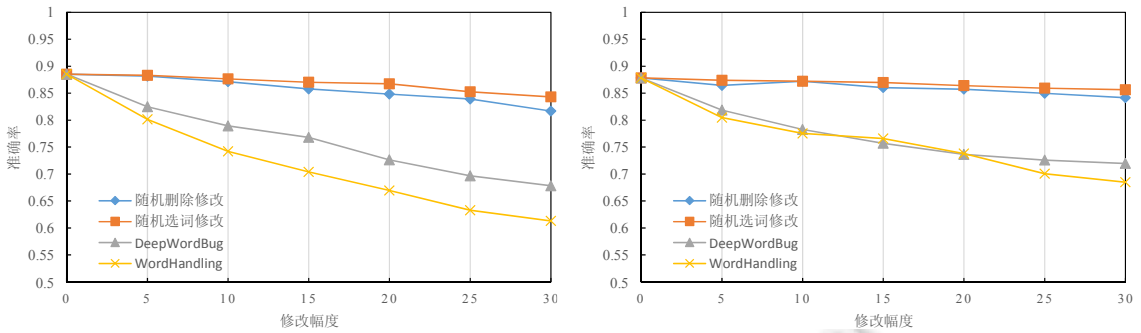
为了验证模型对于对抗样本检测的准确率与样本修改的幅度 m 的关联性,从两种数据集中分别选取 1 000 条长度大于 120 字的数据,根据不同的修改幅度生成相应的对抗样本.图 7 为携程酒店评论数据集在两种模型上检测的准确率随修改规模 m 变化的曲线,图 8 则是在京东购物评论数据集在两种模型上的实验结果.



(a) LSTM 模型对比实验 (b) CNN 模型对比实验

Fig.7 Change rate of accuracy of adversarial examples with the modified amplitude m on Ctrip data

图 7 携程数据对抗样本检测准确率随修改幅度 m 的变化曲线



(a) LSTM 模型对比实验 (b) CNN 模型对比实验

Fig.8 Change rate of accuracy of adversarial examples with the modified amplitude m on Jingdong data

图 8 京东数据抗样本检测准确率随修改幅度 m 的变化曲线

由图 7 和图 8 可以看出,随着修改规模 m 的增大,检测的准确率逐渐降低;即使仅对输入的数据进行个别重要词语的修改,本文提出的 WordHandling 算法也能生成许多对抗样本,误导检测系统的检测.而 m 次修改的总长度最多占输入数据长度的 1/6,超过该数值会严重影响文本的可读性,干扰人对对抗样本内容的理解.

4.3 对抗样本质量度量

图像中距离度量典型的方法是使用 L_p 范数,L_0,L_2,L_∞分别为 3 种常用的 L_p 范数,但其不适用于文本距离度量.因为图像是连续的,而文本是离散的且有词序限制.因此,本文采用 Word Mover's Distance(WMD)[41]对生成的对抗样本质量进行度量.WMD 基于 Earth Mover's Distance(EMD)[42],将 EMD 的适用范围扩展到自然语言处理领域,用于测量两文档之间的距离(即相似性).WMD 距离越大,两文档之间相似性越低;反之则越高.而文档越相似,其语义偏离度则越低.

从生成的对抗样本中随机选取 2 000 条数据进行实验,实验结果如图 9 所示.由该图可看出,WMD 距离小于 0.6 的对抗样本占实验样本总数量的 50%左右,这部分样本与原样本相似度较高;而 WMD 距离大于 0.8 的占样本总数量的 30%左右,这部分样本与原数据相似度较低.

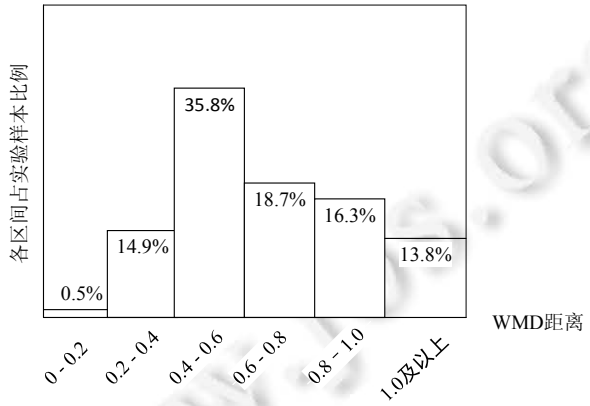


Fig.9 Ratio of sample numbers to total samples in different WMD distance intervals

图 9 不同 WMD 距离区间内样本数量占总样本的比例

对图 9 中各个区间的数据长度进行统计分析,结果见表 4.由该表可以看出,长度大于 30 的文本,在 WMD 距离偏小的区间内所占比例比短文本高.原因在于对抗样本生成过程中修改幅度大小 m,其影响被修改词语在整个输入数据中的比重,短文本数据即使只修改两三字,输入数据也被修改了 10%左右(以长度为 20 字的样本为

例).这也导致与相同修改幅度的长文本相比,短文本的可读性稍差.

Table 4 Proportion of long and short texts in different WMD distance intervals

表 4 不同 WMD 距离区间长短文本数量所占该区间比例

区间	样本长度小于 30(%)	样本长度大于 30(%)
0~0.4	15.4	84.6
0.4~0.6	20.9	79.1
0.6~0.8	24.7	75.3
0.8 及以上	58.1	41.9

表 5 中则给出了几例 WMD 距离计算实例.本文中,与原样本之间 WMD 距离小于 0.6 的对抗样本是语义偏离度较小,阅读性比较好的对抗样本.

Table 5 Examples of WMD distance calculation

表 5 WMD 距离计算实例

样本数据		WMD 距离
原始样本:很一般,性价比很差.跟上海的快捷酒店相比,价格贵,服务差;窗户的高度太低,不小心会摔下去;令人匪夷所思的是,订的是大床房,但是床却是坏的;感觉很不好. 对抗样本:很易班,性价比很岔.跟上海的快捷酒店相比,价格柜,服务岔;窗户的高度泰笛,不小心会摔下去;令人匪夷所思的是,订的是大床房,但是床却是踝的;感觉很部皓.	负面评价 正面评价	0.2196615
原始样本:屏幕较差,拍照也很粗糙. 对抗样本:屏幕交叉,拍照也很出操.	负面评价 正面评价	0.4243181
对抗样本:很易班,性价比很岔.跟上海的快捷酒店相比,价格柜,服务岔;窗户的高度泰笛,不小心会摔下去;令人匪夷所思的是,订的是大床房,但是床却是踝的;感觉很部皓. 对抗样本:屏幕交叉,拍照也很出操.		1.6947902
原始样本:很一般,性价比很差.跟上海的快捷酒店相比,价格贵,服务差;窗户的高度太低,不小心会摔下去;令人匪夷所思的是,订的是大床房,但是床却是坏的;感觉很不好. 原始样本:屏幕较差,拍照也很粗糙.		1.3644687

5 总 结

在本文中,我们提出了中文文本类型的对抗样本生成算法,以此来实现针对网络中深度学习模型的黑盒攻击,诱导这些检测系统做出错误的倾向性判别,使得制作的对抗样本能够避开检测,降低检测的准确率.本文首先利用设计的词语重要性计算函数计算文本数据中的各个词或词组的重要程度,并以此为依据进行排序,针对排在前 m 的词或词组,用同音词替换原词来生成对抗样本,方法有效,且生成的对抗样本内容的改变很小,仍然能够通过上下文或语音谐音来理解语句意思.实验结果表明,本文提出的 WordHandling 算法能够使 LSTM 模型对生成的对抗样本检测的准确率平均降低 29%,使 CNN 模型检测准确率平均降低 22%,且对原始的中文文本的修改幅度仅占输入数据长度的 14.1%.同时,对生成的对抗样本质量进行度量,保证语义偏离度小、可读性好,证明本文提出的 WordHandling 算法有效且表现较佳.此外,文中计算词的重要程度的词语重要性计算函数还能进一步优化,针对每个输入文本,修改幅度 m 的最优选取问题也存在提升的空间.在今后的工作中,我们会对这些存在问题解析并改善提高,并对能够进行定向分类的对抗样本生成进行研究.

References:

[1] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: Proc. of the Advances in Neural Information Processing Systems. 2012. 1097−1105. [doi: 10.1145/3065386]

[2] Taigman Y, Yang M, Ranzato M, Wolf L. Deepface: Closing the gap to human-level performance in face verification. In: Proc. of the Conf. on Computer Vision and Pattern Recognition (CVPR). 2014. 1701−1708. [doi: 10.1109/CVPR.2014.220]

[3] Dahl GE, Yu D, Deng L, Acero A. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. IEEE Trans. on Audio, Speech, and Language Processing, 2012,20(1):30−42. [doi: 10.1109/TASL.2011.2134090]

[4] Collobert R, Weston J. A unified architecture for natural language processing: Deep neural networks with task learning. In: Proc. of the Int'l Conf. on Machine Learning. 2008. 160−167. [doi: 10.1145/1390156.1390177]

[5] Zhang X, Zhao J, Lecun Y. Character-level convolutional networks for text classification. In: Proc. of the Advances in Neural Information Processing Systems. Computer Science, 2015. 649−657. http://arxiv.org/abs/1509.01626v2

[6] Kim Y, Jernite Y, Sontag D, Rush AM. Character-aware neural language models. Association for the Advance of Artificial Intelligence, 2016. https://arxiv.org/pdf/1508.06615v3

[7] Pang B, Lee LL, Vaithyanathan S. Thumbs up? Sentiment classification using machine learning techniques. In: Proc. of the Conf. on Empirical Methods in Natural Language Processing (EMNLP). 2002. 79−86.

[8] Sutskever I, Vinyals O, Le QV. Sequence to sequence learning with neural networks. In: Proc. of the Advances in Neural Information Processing Systems. 2014. 3104−3112. http://arxiv.org/abs/1409.3215v3

[9] Maas AL, Daly RE, Pham PT, Huang D, Ng AY, Potts C. Learning word vectors for sentiment analysis. In: Proc. of the 49th Annual Meeting of the Association for Computational Linguistics. 2011. 142−150.

[10] Kolosnjaji B, Zarras A, Webster G, Eckert C. Deep learning for classification of malware system call sequences. In: Proc. of the Australasian Joint Conf. on Artificial Intelligence. 2016. 137−149. [doi: https://doi.org/10.1007/978-3-319-50127-7_11]

[11] Grosse K, Papernot N, Manoharan P, Backes M, McDaniel P. In: Proc. of the Adversarial Examples for Malware Detection, European Symp. on Research in Computer Security. Cham: Springer-Verlag, 2017. 62−79. [doi: https://doi.org/10.1007/978-3-319-66399-9_4]

[12] Qing SH. Research progress on Android security. Ruan Jian Xue Bao/Journal of Software, 2016,27(1):45−71 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4914.htm [doi: 10.13328/j.cnki.jos.004914]

[13] Rajeswar S, Subramanian S, Dutil F, Pal C, Courville A. Adversarial generation of natural language. In: Proc. of the 2nd Workshop on Representation Learning for NLP. 2017. 241−251. [doi: 10.18653/v1/W17-2629]

[14] Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, Fergus R. Intriguing properties of neural networks. In: Proc. of the Int'l Conf. on Learning Representations (ICLR). 2014.

[15] Ma YK, Wu LF, Jian M, Liu FH, Yang Z. Approach to generate adversarial examples for face-spoofing detection. Ruan Jian Xue Bao/Journal of Software, 2018,29(1):1−10 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5568.htm [doi: 10.13328/j.cnki.jos.005568]

[16] Carlini N, Wagner D. Towards evaluating the robustness of neural networks. In: Proc. of the 2017 IEEE Symp. on Security and Privacy (SP). IEEE, 2017. 39−57. [doi: 10.1109/SP.2017.49]

[17] Liang B, Li H, Su M, Bian P, Li X, Shi W. Deep text classification can be fooled. In: Proc. of the 27th Int'l Joint Conf. on Artificial Intelligence. 2018. 4208−4215. [doi: 10.24963/ijcai.2018/585]

[18] Ebrahimi J, Rao A, Lowd D, Dou D. Hotflip: White-box adversarial examples for text classification. In: Proc. of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018). Melbourne, 2018. https://aclanthology.info/papers/P18-2006/p18-2006

[19] Papernot N, McDaniel P, Swami A, Harang R. Crafting adversarial input sequences for recurrent neural networks. In: Proc. of the Military Communications Conf. (MILCOM 2016). 2016. 49−54.

[20] Papernot N, Mcdaniel P, Goodfellow I, Jha S, Celik ZB, Swami A. Practical black-box attacks against machine learning. In: Proc. of the Asia Conf. on Computer and Communications Security. 2017. [doi: 10.1145/3052973.3053009]

[21] Gao J, Lanchantin J, Soffa ML, Qi Y. Black-box generation of adversarial text sequences to evade deep learning classifiers. In: Proc. of the 2018 IEEE Security and Privacy Workshops (SP Workshops 2018). San Francisco: IEEE, 2018. 50−56.

[22] Barreno M, Nelson B, Sears R, Loseph AD, Tygar AD. Can machine learning be secure? In: Proc. of the ACM Symp. on Information, Computer and Communications Security. ACM Press, 2006. 16−25. [doi: 10.1145/1128817.1128824]

[23] Rubinstein BIP, Nelson B, Huang L, Joseph AD, Lau S, Rao S, Taft N, Tygar JD. Antidote: Understanding and defending against poisoning of anomaly detectors. In: Proc. of the 9th ACM SIGCOMM Conf. on Internet Measurement Conf. ACM Press, 2009. 1−14. [doi: 10.1145/1644893.1644895]

[24] Shafahi A, Huang WR, Najibi M, Suciu O, Studer C, Dumitras T, Goldstein T. Poison frogs! Targeted clean-label poisoning attacks on neural networks. In: Proc. of the Advances in Neural Information Processing Systems. 2018. No.7849.

[25] Biggio B, Corona I, Maiorca D, Nelson B, Šrndić N, Laskov P, Giacinto G, Roli F. Evasion attacks against machine learning at test time. In: Proc. of the Joint European Conf. on Machine Learning and Knowledge Discovery in Databases. Springer-Verlag, 2013. 387−402. [doi: 10.1007/978-3-642-40994-3_25]

[26] Šrndić N, Laskov P. Practical evasion of a learning-based classifier: A case study. In: Proc. of the 2014 IEEE Symp. on Security and Privacy. Washington: IEEE Computer Society, 2014. 197−211. [doi: 10.1109/SP.2014.20]

[27] Liang B, Su M, You W, Shi W, Yang G. Cracking classifiers for evasion: A case study on the Google's phishing pages filter. In: Proc. of the 25th Int'l Conf. on World Wide Web. 2016. 345−356. [doi: 10.1145/2872427.2883060]

[28] Goodfellow IJ, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. In: Proc. of the Int'l Conf. on Learning Representations. 2015.

[29] Kereliuk C, Sturm B, Larsen J. Deep learning and music adversaries. IEEE Trans. on Multimedia, 2015,17(11):2059−2071. [doi: 10.1109/TMM. 2015.2478068]

[30] Nguyen A, Yosinski J, Clune J. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In: Proc. of the 2015 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). IEEE, 2015. [doi: 10.1109/CVPR.2015.7298640]

[31] Papernot N, Mcdaniel P, Jha S, Fredrikson M, Celik ZB, Swami A. The limitations of deep learning in adversarial settings. In: Proc. of the 2016 IEEE European Symp. on Security and Privacy (EuroS&P). IEEE, 2016. 372−387. [doi: 10.1109/EuroSP.2016.36]

[32] Moosavidezfooli SM, Fawzi A, Frossard P. DeepFool: A simple and accurate method to fool deep neural networks. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). IEEE, 2016. [doi: 10.1109/CVPR.2016.282]

[33] Johnson R, Zhang T. Effective use of word order for text categorization with convolutional neural networks. In: Proc. of the 2015 Annual Conf. of the North American Chapter of the ACL. 2015. 103−112. [doi: 10.3115/v1/N15-1011]

[34] Johnson R, Zhang T. Supervised and semi-supervised text categorization using LSTM for region embeddings. In: Proc. of the Int'l Conf. on Machine Learning. 2016. 526−534.

[35] Lecun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proc. of the IEEE, 1998,86(11): 2278−2324. [doi: 10.1109/5.726791]

[36] Kim Y. Convolutional neural networks for sentence classification. In: Proc. of the Conf. on Empirical Methods in Natural Language Processing (EMNLP). 2014. 1746−1751. [doi: 10.3115/v1/D14-1181]

[37] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997,9(8):1735−1780. [doi: 10.1162/neco.1997.9.8.1735]

[38] Takeru M, Dai Andrew M, Ian G. Adversarial training methods for semi-supervised text classification. In: Proc. of the Int'l Conf. on Learning Representations. 2017.

[39] Sundermeyer M, Ney H, Schluter R. From feedforward to recurrent LSTM neural networks for language modeling. IEEE/ACM Trans. on Audio, Speech, and Language Processing, 2015,23(3):517−529. [doi: 10.1109/TASLP.2015.2400218]

[40] Graves A, Jaitly N, Mohamed AR. Hybrid speech recognition with deep bidirectional LSTM. In: Proc. of the Automatic Speech Recognition and Understanding. IEEE, 2014. 273−278. [doi: 10.1109/ASRU.2013.6707742]

[41] Kusner MJ, Sun Y, Kolkin NI, Weinberger KQ. From word embeddings to document distances. In: Proc. of the Int'l Conf. on Int'l Conf. on Machine Learning. 2015. 957−966.

[42] Rubner Y, Tomasi C, Guibas LJ. The earth mover's distance as a metric for image retrieval. Int'l Journal of Computer Vision, 2000, 40(2):99−121. [doi: 10.1023/A:1026543900054]

附中文参考文献:

[12] 卿斯汉.Android 安全研究进展.软件学报,2016,27(01):45−71. http://www.jos.org.cn/1000-9825/4914.htm [doi: 10.13328/j.cnki.jos.004914]

[15] 马玉琨,毋立芳,简萌,刘方昊,杨洲.一种面向人脸活体检测的对抗样本生成算法.软件学报,2018,29(1):1−10. http://www.jos.org.cn/1000-9825/5568.htm [doi: 10.13328/j.cnki.jos.005568]

王文琦(1992－),男,湖北襄阳人,博士生,主要研究领域为人工智能安全,自然语言处理.

王丽娜(1964－),女,博士,教授,博士生导师,主要研究领域为系统安全,信息隐藏.

汪润(1991－),男,博士,主要研究领域为移动设备隐私保护,机器学习.

唐奔宵(1991－),男,博士,CCF 学生会员,主要研究领域为 Android 隐私保护,机器学习.