计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200500078

面向自然语言处理的深度学习对抗样本综述

仝 鑫 王斌君 王润正 潘孝勤
中国人民公安大学信息网络安全学院 北京 100038
(tongxindotnet@outlook.com)

摘 要 深度学习模型被证明存在脆弱性并容易遭到对抗样本的攻击,但目前对于对抗样本的研究主要集中在计算机视觉领域而忽略了自然语言处理模型的安全问题。针对自然语言处理领域同样面临对抗样本的风险,在阐明对抗样本相关概念的基础上,文中首先对基于深度学习的自然语言处理模型的复杂结构、难以探知的训练过程和朴素的基本原理等脆弱性成因进行分析,进一步阐述了文本对抗样本的特点、分类和评价指标,并对该领域对抗技术涉及到的典型任务和数据集进行了阐述;然后按照扰动级别对主流的字、词、句和多级扰动组合的文本对抗样本生成技术进行了梳理,并对相关防御方法进行了归纳总结;最后对目前自然语言处理对抗样本领域攻防双方存在的痛点问题进行了进一步的讨论和展望。

关键词:自然语言处理;深度学习;人工智能安全;对抗样本;鲁棒性
中图法分类号 TP301

Survey on Adversarial Sample of Deep Learning Towards Natural Language Processing

TONG Xin, WANG Bin-jun, WANG Run-zheng and PAN Xiao-qin
School of Information and Cyber Security, People's Public Security University of China, Beijing 100038, China

Abstract Deep learning models have been proven to be vulnerable and easy to be attacked by adversarial examples, but the current researches on adversarial samples mainly focus on the field of computer vision and ignore the security of natural language processing models. In response to the same risk of adversarial samples faced in the field of natural language processing(NLP), this paper clarifies the concepts related to adversarial samples as the basis of further research. Firstly, it analyzes causes of vulnerabilities, including complex structure of the natural language processing model based on deep learning, the training process that is difficult to detect and the naive basic principles, further elaborates the characteristics, classification and evaluation metrics of text adversarial examples, and introduces the typical tasks and classical datasets involved in the adversarial examples related to researches in the field of natural language processing. Secondly, according to different perturbation levels, it sorts out various text adversarial examples generation technology of mainstream char-level, word-level, sentence-level and multi-level. What's more, it summarizes defense methods, which are relevant to data, models and inference, and compares their advantages and disadvantages. Finally, the pain points of both attack and defense sides in the field of current NLP adversarial samples are further discussed and anticipated.

Keywords Natural language processing, Deep learning, AI security, Adversarial examples, Robustness

1 引言

自然语言处理(Natural Language Processing,NLP)技术在深度学习的推动下正蓬勃发展,并在机器翻译、文本分类等方面超越了基于统计机器学习的方法。一方面,LSTM[1]等循环神经网络模型通过与 Word2vec[2],Glove[3]等词嵌入模型相互配合,能够挖掘到文本时序层次的信息,避免了传统机器学习模型的特征损失;另一方面,以 BERT[4],XLNet[5]为代表的预训练语言模型通过使用海量的数据集和无监督的训练方法,具备深层网络结构且规模更大,进一步解决了一词多义等难题,并在多个 NLP 任务中拥有媲美人类的表现,成为 NLP 领域中新的里程碑。

遗憾的是,神经网络算法自身固有的局部线性以及数据的高维度等特点导致基于深度学习的 NLP 模型的安全性受限[6],容易受到对抗和欺骗。自 Szegedy 等[7]首次发现对图像添加微小扰动可误导神经网络做出错误分类后,大批研究

到稿日期:2020-05-18 返修日期:2020-08-25 本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:2020 CCF-绿盟科技"鲲鹏"科研基金(CCF-NSFOCUS 2020011);公安部科技强警基础专项(2018GABJC03);国家社会科学基金重点项目(20AZD114);中国人民公安大学拔尖创新人才培养经费支持硕士研究生项目(2020ssky005);中国人民公安大学公共安全行为科学研究与技术创新项目
This work was supported by the 2020 CCF-Nsfocus "Kunpeng" Research Fund(CCF-NSFOCUS 2020011), Science and Technology Strengthening Police Basic Program of Ministry of Public Security(2018GABJC03), Key Program of the National Social Science Foundation of China (20AZD114), Top Talent Training Special Funding Graduate Research and Innovation Project of People's Public Security University of China (2020ssky005), and Scientific Research and Technological Innovation on Public Security Behavior of People's Public Security University of China.
通信作者:王斌君(wangbinjun@ppsuc.edu.cn)

者针对计算机视觉模型开展对抗样本及其防御方法的研究,并提出了一系列经典的对抗样本生成方法[8]。但是,目前针对 NLP 领域的对抗方法的探索却寥寥无几,Wang 等[6]的统计分析表明:2014 年出现了对抗样本的研究,2016 年出现第一篇文本领域的对抗样本研究,到 2019 年对抗样本研究文献总计 696 篇,文本领域研究 21 篇。无论从绝对数还是该领域的增长率来看,文本对抗样本的研究尚处于萌芽阶段。

基于深度学习的 NLP 技术已经在日常生产、生活的各个方面得到越来越广泛的应用,其鲁棒性和安全性问题更不可忽视。因此,为了使研究者能够全面了解文本对抗研究领域的现状,本文主要从对抗样本的概念、NLP 模型脆弱性的成因、文本对抗样本的评估指标、常用公开数据集、生成及防御方法等方面进行综述,讨论了目前该领域面临的挑战,并对文本对抗样本的未来发展进行了预测和展望。

2　NLP 领域的对抗样本

2.1　对抗样本的基本概念

对抗样本是指对输入数据 $\boldsymbol{x}$ 添加人眼难以察觉的微小扰动而构造成的样本 $\boldsymbol{x}'$,当 $\boldsymbol{x}'$ 输入到训练好的深度学习模型时,模型将以较高的置信度识别出与原始标签不同的输出结果,如式(1)所示:

$$f(\boldsymbol{x})\neq f(\boldsymbol{x}') \quad \text{s.t.} \quad \|\boldsymbol{x}'-\boldsymbol{x}\|<\boldsymbol{\varepsilon} \tag{1}$$

其中,f 表示神经网络的正向传播过程,$\boldsymbol{\varepsilon}$ 用于衡量和限制扰动的大小。

2.2　NLP 模型缺陷及其对抗样本的特点

2.2.1　NLP 深度学习模型的脆弱性分析

从模型复杂度来看,基于深度学习的 NLP 模型已经取代传统的统计学习模型,受 transformer[9] 启发的、基于编码器-解码器(encoder-decoder)结构的大规模预训练语言模型更是成为主流,其包含注意力机制、跳跃连接等复杂结构,网络深度能达数百层。尽管模型复杂度的增加能够进一步提高其表征能力,但同时也使其愈发近似于“黑盒”模型以至难以理解和调试。

从训练方法来看,目前 NLP 模型主要采用端到端的训练方式,即输入的是原始文本(始端),输出的是最终目标(末端),中间过程难以解释和认知,这导致模型在受到对抗攻击时无法准确定位具体漏洞的成因并制定合理的修复措施。

从基本原理来看,语言模型始终是 NLP 任务的核心,无论是传统的隐马尔可夫模型、n 元模型还是 LSTM,BERT 等基于深度学习的语言模型,其基本原理都是利用文本元素的联合概率分布来衡量该文本存在的可能性,如式(2)所示:

$$P(\boldsymbol{w}_1,\cdots,\boldsymbol{w}_m)=P(\boldsymbol{w}_1)P(\boldsymbol{w}_1\mid\boldsymbol{w}_2)\cdots P(\boldsymbol{w}_m\mid\boldsymbol{w}_1,\cdots,\boldsymbol{w}_{m-1}) \tag{2}$$

这些模型关注的仍然是样本的字、词、句的低层统计特征甚至虚假统计特征[10],很难挖掘到语义级别的高层次抽象特征,使得对关键词进行修改、删除或增加等操作会影响到模型最终的输出概率分布,从而导致出现对抗样本攻击。

2.2.2　NLP 深度学习模型的对抗样本特点

文本数据和图像数据存在显著的差异性,因此面向 NLP 模型的对抗样本无论是生成方法还是对抗目标都与面向计算机视觉的对抗样本有很大不同[6]。一方面,图像数据的像素具有连续性特征,使得利用优化或梯度的方法对像素值进行微小的改变来生成对抗样本成为可能。而文本数据的离散属性导致直接利用梯度方法为输入数据增加噪声会生成与自然语言语法相悖、语义难以为人所理解的样本。另一方面,在文本的对抗攻击中,小的扰动如拼音、错别字或形近字等虽然很容易被察觉,但人们仍然能根据上下文“猜出”文本想要表达之意。因此,文本对抗样本更关注如何欺骗深度学习系统,而未必需要严格追求生成人类完全无法分辨的扰动。

这些差异不仅导致用于图像的对抗攻击方法无法直接应用于文本数据,也使得 NLP 模型在防御对抗样本攻击时,除了要研究人类不可察觉的特征鲁棒外,还要关注人类可辨识的特征鲁棒,因此更具挑战性。

2.2.3　NLP 深度学习模型的对抗样本分类

如图 1 所示,主流的对抗样本分类标准主要以已知的信息水平为依据,可划分为:

(1)黑盒攻击。对模型、训练集不了解或了解很少,仅能观测到模型的输入和输出结果,而观测不到模型内部的参数和梯度信息。

(2)白盒攻击。对模型和训练集完全了解,并且能够观测到模型在前向传播和反向传播时内部梯度和参数的变化。

针对攻击时是否可选择希望模型错分的类别,可以将其划分为:

(1)定向攻击。使样本被错分到指定的类别。

(2)非定向攻击。只需导致模型分类出错,而不预先指定错分的类别。

此外,NLP 模型通常使用字符编码或词编码作为模型输入特征,因此又可依据针对这些特征的扰动级别将文本对抗样本划分为:

(1)字符级对抗样本。通过修改文本中的字符,如英文字母或中文汉字来构建的对抗样本。

(2)词组级对抗样本。对原始文本中的关键词进行删除、替换或插入新词。

(3)语句级对抗样本。为原始文本添加精心构造的句子以改变模型识别结果。

(4)多级对抗样本。混合使用了上述 3 种扰动方法生成的对抗样本。

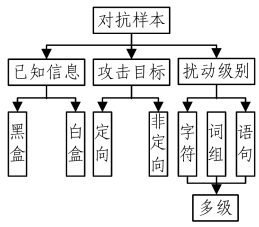

图 1　对抗样本的分类

Fig. 1　Classification of adversarial examples

3　NLP 领域的对抗样本评估

从上述对抗样本的定义及特点可知,对抗样本攻击的本质是在保持自身隐蔽性的同时尽可能地误导和修改模型的输出结果,这就决定了对对抗样本的评估涵盖两个层次:对抗样本本身隐蔽性的评价指标和针对目标模型的破坏程度。

3.1　对抗样本的评价指标

对抗样本本身的性质是指对抗样本的不可察觉性或扰动

幅度。图像领域利用 $L0$，$L2$ 和 $L\infty$ 范数来表示对抗样本与原始样本相似程度的方法并不适用于文本，文本领域主要利用能够衡量语句真实性和扰动幅度的文本相似性度量作为指标，具体如下。

(1)困惑度。计算语言模型生成句子 $\boldsymbol{w}$ 质量的好坏，其值越小表示文本越贴近于人的表达方式，如式(3)所示：

$$PP(\boldsymbol{w})=2^{-\frac{1}{N}\sum_{i=1}^{N}\log(p(\boldsymbol{w}_i))}=\sqrt[N]{\prod_{i=1}^{N}\frac{1}{p(\boldsymbol{w}_i\mid \boldsymbol{w}_1,\boldsymbol{w}_2,\cdots,\boldsymbol{w}_{i-1})}} \tag{3}$$

(2)余弦相似度。衡量文本向量 $\boldsymbol{p}$ 和 $\boldsymbol{q}$ 之间的相似性，其值越小表示对抗样本越接近真实样本，如式(4)所示：

$$\cos(\boldsymbol{p},\boldsymbol{q})=\frac{\sum_{i=0}^{N}p_i\times q_i}{\sqrt{\sum_{i=0}^{N}(p_i)^2}\times\sqrt{\sum_{i=0}^{N}(q_i)^2}} \tag{4}$$

(3)词移距离[11]。表示句子向量 $\boldsymbol{p}$ 中所有的词汇 p_i 移动到句子向量 $\boldsymbol{q}$ 中词汇 q_j 的位置的距离总和，如式(5)所示。其中，$T_{i,j}$ 代表当前计算词汇的距离权重，d 代表文本欧氏距离。

$$WMD(\boldsymbol{p},\boldsymbol{q})=\sum_{i,j=0}^{M,N}T_{i,j}\times d(p_i,q_j) \tag{5}$$

改进版的词移距离[12]和词移动嵌入[13]方法进一步提升了 WMD 在文本分类问题上的表现和计算效率，能更好地度量文本对抗样本的扰动噪声大小。

(4)杰卡德相似系数。用于衡量句子向量 $\boldsymbol{p}$ 和 $\boldsymbol{q}$ 之间的相同词汇比例，如式(6)所示：

$$J(\boldsymbol{p},\boldsymbol{q})=\frac{|\boldsymbol{p}\cap\boldsymbol{q}|}{|\boldsymbol{p}\cup\boldsymbol{q}|} \tag{6}$$

3.2 目标模型的评价指标

目标模型的表现是指遭受对抗攻击或采取防御措施后模型本身性能的变化。除了准确率、混淆矩阵和 AUC(Area Under Curve)面积等传统评价模型效果的指标外，一些衡量对抗样本攻击破坏程度的指标被相继提出。

(1)攻击成功率。它指在对抗样本的攻击下，导致模型分类出错的样本比例。对于定向攻击，在对抗样本生成算法 A 攻击下的攻击成功率如式(7)所示：

$$Asr=\frac{1}{N}\sum_{i=1}^{N}I(f(A(\boldsymbol{x}_i))=y_i^*) \tag{7}$$

其中，y_i^* 是定向攻击的目标类别。对于非定向攻击，其攻击成功率只需考虑分类结果与样本原始类别 y_i 不同的情况，如式(8)所示：

$$Asr=\frac{1}{N}\sum_{i=1}^{N}I(f(A(\boldsymbol{x}_i))\neq y_i) \tag{8}$$

(2)扰动变化曲线和攻击强度曲线。Dong 等[14]提出，通过观察目标模型的分类准确率或攻击算法的攻击成功率随扰动程度 $\boldsymbol{\varepsilon}$ 和攻击迭代次数等强度指标变化的曲线，能够动态化地刻画攻击效果。

特别地，Michel 等[15]同时考虑了对抗样本和目标模型两方面的评估指标，为对抗样本增加了更加严格的约束。在增加扰动前后，对抗样本和原始样本应具备较近的语义距离，而具备较远的向量空间距离以实现误导模型输出错误结果。对比实验证明，chrF 评估指标相比 BLEU 等更贴近人类的表现。在此基础上本文针对 seq2seq 文本模型提出了能够判断增加扰动前后样本语义等价性的对抗攻击评估框架，从而更客观、全面地评估对抗样本的攻击及防御措施的效果。

4 对抗样本生成方法

4.1 实验常用数据集概览

表 1 列出了面向 NLP 系统的在对抗样本实验中常用的部分经典数据集，以便读者更加直观、全面地了解下文所述对抗方法的任务目标和攻击效果。

表 1 NLP 对抗样本研究常用数据集

Table 1 Common datasets in NLP adversarial example researches

任务	数据集	规模/K	链接	描述
TC	IMDB	50(2 classes)	http://ai.stanford.edu/~amaas/data/sentiment/	包含正负情感鲜明的电影评论数据，并额外提供了无标注数据
TC	AG news[16]	20(4 classes)	http://www.di.unipi.it/~gulli/AG_corpus_of_news_articles.html	从超过 2000 个新闻源收集的数据集，可用于新闻分类，提供了 db 和 xml 两种格式版本
TC	SST[17]	12(5 classes)	https://nlp.stanford.edu/sentiment/treebank.html	来自斯坦福的标准情感数据集，其中每个句子分析树的节点均有细粒度的情感注解
TC	Tan Songbo's Hotel Reviews	6(2 classes)	https://www.aitechclub.com/data-detail?data_id=29	数据采集自携程酒店评论，文本长度偏短，标签均衡
MT	TED (WIT3)[18]	en-fr:238, en-de:219, en-zh:238	https://wit3.fbk.eu/	所收文本源自 TED.com 网站演讲视频，包含英文与多种语言的翻译数据
QA	SQuAD[19-20]	v1.1:train:87 599, val:10 570, test:9 533, v2.0:train:130, val:11 873, test:8 862	https://rajpurkar.github.io/SQuAD-explorer/	数据来源于维基百科，问题的答案均包含于对应阅读段落中，2.0 版本包含无解问题，可进一步检验模型回答的有效性
QA	Visual QA[21]	Images:200, question:1 000	https://visualqa.org/	题目包含图像和文本两种数据类型，可训练计算机视觉与 NLP 相结合的综合问答系统
QA	WebQA	train:36 181, val:3 018; test:3 024	http://idl.baidu.com/WebQA.html	基于百度知道和其他资源构建的中文问答数据集
TE	SNLI[22]	sentence pairs:570	https://nlp.stanford.edu/projects/snli/	人工标注了平衡的分类标签：蕴含、矛盾、中性 3 种
TE	MNLI[23]	sentence pairs:430	https://www.nyu.edu/projects/bowman/multinli/	类似于 SNLI，同时覆盖了多种口头和书面文本的语型
TE	Chinese Textual Inference	sentence pairs:880	https://github.com/liuhuanyong/ChineseTextualInference	用于构建中文文本推断模型的大型数据集，并提供了基于 Bi-LSTM 的基准参考结果
NER	CoNLL 2003[24]	en:1393; de:909	https://www.clips.uantwerpen.be/conll2003/ner/	包含人、地、机构和其他实体 4 类标签
NER	CLUENER 2020	12	https://github.com/CLUEbenchmark/CLUENER2020	基于 THUCNEWS 数据标注得到的细粒度中文 NER 数据集，包含 10 类实体标签

注：TC 表示文本分类(Text Classification)，NER 表示命名实体识别(Named Entity Recognition)，MT 表示机器翻译(Machine Translation)，QA 表示机器阅读/问答系统(Question & Answer)，TE 表示文本蕴含(Text Entailment)

文本分类数据集通常以短文本语料为主,主要用于字符级、词组级等小扰动幅度的对抗实验,通用的效果衡量指标是准确率降幅和攻击成功率。机器翻译、问答、文本蕴含等包含长文本较多的数据集还广泛用于语句级的对抗文本生成实验,除了采用错误率、F1 值等评价指标,还引入了 BELU 和人类参与测试的方法以考查生成文本的自然程度和语法语义错误。同时考虑到针对中文领域和命名实体识别任务的研究较少,额外补充了部分相关的公开数据集,以用于后续研究工作。所有数据集均提供了下载链接,以便读者复现和开展 NLP 对抗样本的相关实验。

4.2　字符级对抗样本

人类具有鲁棒性很强的自然语言理解能力,即使遇到包含错漏字的文本也不会影响阅读和理解。然而相比人类,NLP 系统则脆弱得多。2017 年,Facebook 的神经机器翻译系统曾将原本表达“早上好”的阿拉伯语理解成“攻击他们”,原因是其包含了一个错误字母(二者阿拉伯语的差异如图 2 所示),最终导致一名无辜群众遭受牢狱之灾。这表明通过对文本字符、数字甚至标点符号的微小调整,能够显著地影响基于深度学习的 NLP 模型的输出结果。这个事件引起了研究者们对字符级文本对抗样本的关注。

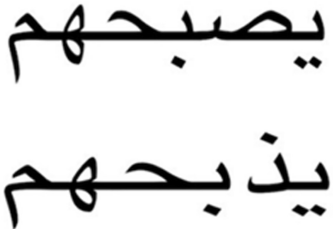

图 2　阿拉伯语“早上好”(上)和“攻击他们”(下)

Fig. 2　“Good morning”(above) and “attack them” (below) in Arabic

字符级对抗样本的研究重点在于如何精确地定位对预测结果有显著影响的关键字符,从而实现在付出最小扰动代价的情况下生成不易察觉的对抗样本。相关探索经历了“随机扰动”到“基于评分的黑盒方法”再到“基于梯度的白盒攻击”3 个发展阶段。

文本字符的离散性特征为使用基于微分的白盒对抗样本生成方法带来了困难,早期的字符级对抗样本生成方法主要采用黑盒攻击的形式。Belinkov 等[25]利用天然扰动和人工合成扰动两种方式搜集和生成对抗样本。他们搜集了一批包含字符错误的语料来代表天然扰动对抗样本,即人在非故意情况下出现的语法或拼写失误,并进一步采用单词内字母顺序交换、字母随机替换和键盘位邻近字母互换等方式人工模拟可能的字符级错误以生成合成扰动数据集。最终在 TED 演讲数据集上针对基于 char2char,Nematus 和 char-CNN 的机器翻译系统发起攻击,并导致上述模型翻译结果的 BLEU 值显著下降,且基于天然扰动的对抗样本的影响力更强。但是该方法仅通过随机字符修改的方式验证了对抗样本的存在性,而没有讨论如何精确定位关键字符。笔者在中文微博情感分类数据集上的复现实验中发现,上述基于随机修改的方法在扰动比例高达 85%时产生的对抗文本效果仍然逊于扰动比例仅为 5%的包含精确定位机制的方法。

后续的探索主要围绕如何高效地定位关键扰动点以增强对抗隐蔽性展开,Gao 等[26]在 Belinkov 实验的基础上增加了关键词定位技术,提出在黑盒条件下生成包含精确、轻微扰动的对抗样本方法 DeepWordBug,具体包含以下两步。

(1)时序评分。通过观察去掉语句中的某个单词 $\boldsymbol{w}_i$ 后模型的表现来衡量该单词的重要程度,包括首部重要性 $THS(\boldsymbol{w}_i)$ 和尾部重要性 $TTS(\boldsymbol{w}_i)$,如式(9)和式(10)所示:

$$THS(\boldsymbol{w}_i)=f(\boldsymbol{w}_1,\cdots,\boldsymbol{w}_{i-1},\boldsymbol{w}_i)-f(\boldsymbol{w}_1,\cdots,\boldsymbol{w}_{i-1}) \tag{9}$$

$$TTS(\boldsymbol{w}_i)=f(\boldsymbol{w}_i,\boldsymbol{w}_{i+1},\cdots,\boldsymbol{w}_{end})-f(\boldsymbol{w}_{i+1},\cdots,\boldsymbol{w}_{end}) \tag{10}$$

然后,通过超参数 λ 平衡二者得到最终的联合评分结果,如式(11)所示:

$$CombinedScore(\boldsymbol{w}_i)=THS(\boldsymbol{w}_i)+\lambda\times TTS(\boldsymbol{w}_i) \tag{11}$$

(2)改变字符。利用与文献[25]人工合成扰动类似的增删、交换方法来修改字符,以生成对抗样本。

DeepWordBug 在 AG News 等经典分类任务数据集上使得 word-LSTM 和 char-CNN 模型的准确率分别下降了 68%和 48%,尽管这种时序评分机制能够在模型结构和梯度信息未知的情况下简单、高效地定位扰动关键点,并影响后续一系列的黑盒方法,但这种将句子分裂后进行二次处理的方法可能对句子的完整性产生负面影响并导致信息损失。Wang 等[27]将时序评分拆开的两部分进行合并,提出了词删除评分机制,用于衡量目标中文字符的重要性,如式(12)所示:

$$DS(\boldsymbol{w}_i)=f(\boldsymbol{w}_1,\cdots,\boldsymbol{w}_{i-1},\boldsymbol{w}_i,\boldsymbol{w}_{i+1},\cdots,\boldsymbol{w}_n)-f(\boldsymbol{w}_1,\cdots,\boldsymbol{w}_{i-1},\boldsymbol{w}_{i+1},\cdots,\boldsymbol{w}_n) \tag{12}$$

最终构建了基于中文同音字替换的黑盒对抗样本生成方法 WordHandling,这也是目前为止为数不多的面向中文 NLP 模型的对抗样本构造方法之一。然而,评分机制并非是完美无缺的。一方面,单词重要性会随着语句长度的增长而逐渐分散,受程序语言表达浮点数不精确的限制,评分机制无法精确察觉到细粒度的分数变化;另一方面,这种机制的主要依据是准确率的变化,因此可被广泛应用于面向分类的 NLP 模型,但不适用于机器翻译等模型。

在模型结构和梯度信息已知的情况下,白盒攻击方法能够跟踪输出结果到输入文本序列反向传播时的梯度信息,相比黑盒方法能够观察到更加细微的改变,从而实现关键词的高精准定位。Ebrahimi 等[28]尝试应对文本数据离散性带来的挑战,并在探索白盒方法上取得了进展,提出了针对机器翻译的白盒攻击方法。该方法首先对文本使用字符级 one-hot 编码,定义词 a 转换为词 b 时,第 j 个字母的位置交换向量如式(13)所示:

$$\vec{v}_{ijb}=(\vec{0},\cdots;(\vec{0},\cdots,-1,0,\cdots,1,0)_j;\cdots,\vec{0})_i;\vec{0},\cdots) \tag{13}$$

然后利用翻译源序列 $\boldsymbol{x}$ 和目标序列 $\boldsymbol{y}$ 之间的对数损失 $J(\boldsymbol{x},\boldsymbol{y})$ 对 $\vec{v}_{ijb}$ 的微分,来近似估计 a 单词通过改变字符变为 b

单词后的损失变化，如式(14)所示：

$$\nabla_{\vec{v}_{ijb}} J(\boldsymbol{x},\boldsymbol{y}) = \nabla_x J(\boldsymbol{x},\boldsymbol{y})^{\mathrm{T}} \cdot \vec{\boldsymbol{v}}_{ijb} = \frac{\partial J^{(b)}}{\partial \boldsymbol{x}_{ij}} - \frac{\partial J^{(a)}}{\partial \boldsymbol{x}_{ij}} \tag{14}$$

最后配合贪心搜索和束搜索方法，进一步寻找最佳的扰动变化来完成定向或非定向攻击。在对抗基于 TED 数据集训练的神经机器翻译模型的实验中其效果优于 Belinkov 等[25]提出的黑盒方法。

此外，VIPER[29]方法的思路不同于上述通过定位关键词来减少扰动幅度并最终达到增强隐蔽性的方法，而是另辟蹊径，通过强调字符扰动时的视觉近似性来减少人类察觉的可能性，使得替换后的效果相比其他基于 26 个英文字母进行替换的方法更难以被发现，兼顾了隐蔽性和攻击效果，能够导致部分 NLP 模型性能下降 82%，其效果如图 3 所示(图片出自文献[29])。

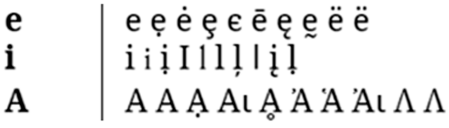

图 3 VIPER 的对抗样本更具视觉近似性

Fig. 3 Adversarial examples generated by VIPER have better visual similarity

4.3 词组级对抗样本

虽然字符级 NLP 模型可不受词嵌入库规模的限制而识别更多的词汇，但 Word2vec，BERT(英文版本)等词特征向量构建技术能够发现近义词、反义词等更高层次的语义信息，这是字符级模型无法比拟的，因此词组级模型仍是 NLP 的主流技术。这也使得通过增删和替换关键词的方式来生成对抗样本从而影响模型结果成为可能。

对于词级文本对抗样本，首要问题仍然是关键词的精准定位。一些实验采用了与字符级方法相仿的梯度计算和评分的机制，如 Papernot 等[30]受图像对抗样本技术的启发，尝试在文本领域复现视觉领域的白盒对抗样本生成技术。他们提出通过计算 RNN 模型雅克比矩阵的方式来量化输入序列第 i 时间步的特征 $\boldsymbol{x}^{(i)}$ 对输出序列第 j 时间步特征 $\boldsymbol{y}^{(j)}$ 的贡献，如式(15)所示，其原理与字符级研究[28]相似，主要区别在于后者额外考虑到了改变字符前后梯度的改变量。

$$J_f[i,j] = \frac{\partial \boldsymbol{y}^{(j)}}{\partial \boldsymbol{x}^{(i)}} \tag{15}$$

基于此，文献[30]成功将图像领域常用的快速梯度下降(Fast Gradient Sign Method，FGSM)[31]对抗样本生成方法用于修改对结果影响最显著的词，其针对 RNN 生成的对抗序列能在平均修改语句中 13.1%单词的情况下导致文本分类模型 100%出错。Jin 等[32]提出在 NLP 领域逐渐取代 RNN 的 BERT 模型的黑盒对抗攻击方法 TEXTFOOLER。该方法先利用类似文献[26]的词级时序评分机制对单词重要性进行排序；然后对关键词进行同义词替换，并选取替换前后余弦相似度较大且显著改变模型结果的样本作为对抗样本。在攻击用于 IMDB 分类的 BERT 模型时，其成功率高达 99.7%。

另一些研究通过设计更丰富的特征来综合考查词组的重要程度，Samanta 等[33]尝试寻找与特定任务相关的词语并进行修改。例如，在电影评论分类任务中，除了一些全局、通用的情感词如“good(好)”“bad(坏)”能够表达文本的感情倾向外，“hilarious(搞笑)”一词对于喜剧片可能是褒义，而对于恐怖电影可能就是讽刺，因此通过修改这种与特定任务相关的文本同样能够实现对抗攻击。这种方法的优点在于提升了攻击搜索空间的大小，使得攻击更容易成功，但缺点是牺牲了对抗样本的迁移性。

字符级对抗样本的生成思路主要是改变单词中不起眼的字母，使单词在预测时“失效”，其攻击策略的侧重点在于提升隐蔽性且无须考虑语法语义层次问题。词组级对抗样本修改的是完整的单词，很难做到不被察觉，而一些对抗方法虽然能够对多种 NLP 模型发起卓有成效的攻击，但其扰动算法缺乏可解释性，可能无法将词嵌入空间的扰动还原为真实存在的单词，并会导致生成的对抗样本出现语法、语义错误等。因此，需要考虑生成样本的真实性和流畅性以避免引起阅读者的注意，主要的思路有 4 种：1)增加扰动约束以消除扰动词向量的异常映射现象；2)在训练过程中引入衡量样本流畅性的指标；3)使用辅助算法；4)配合外部知识库实施攻击策略。

Sato 等[34]提出了一种对词嵌入空间中单词的扰动方向进行约束的方法 iAdvT-Text，如图 4 所示。将单词扰动的搜索空间限制在一个线性空间中，使得添加扰动后生成的对抗样本可以对应具体的自然语言，从而减少拼写或语法错误。

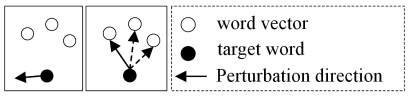

图 4 无扰动方向限制(左)和限制扰动方向(右)

Fig. 4 Perturbation direction without limitation(left) or with limitation(right)

Zhang 等[35]采用基于 Metropolis-Hastings 抽样的 MHA 方法来提升对抗样本的流畅性。该算法包含黑盒攻击版本 b-MHA 和白盒攻击版本 w-MHA，并设计了关于对抗样本的平稳分布，如式(16)所示：

$$\pi(\boldsymbol{x}|y^*) \propto LM(\boldsymbol{x}) \cdot f(y^*|\boldsymbol{x}) \tag{16}$$

其中，$LM(\boldsymbol{x})$为语言模型 LM 生成句子 $\boldsymbol{x}$ 的概率，用于衡量样本流畅性；$f(y^*|\boldsymbol{x})$表示句子被错误识别为目标标签 y^* 的概率。该方法最终能够生成流畅度较高的对抗样本，但性能受到抽样准确率的限制。

Alzantot[36]通过使用遗传算法解决了离散文本的生成和搜索难题，以提升对抗样本质量。92.3%的对抗样本甚至能够欺骗人类受试者，但是遗传算法收敛慢的缺点在一定程度上增加了算法的攻击代价。Zang 等[37]采用粒子群算法辅助文本修改工作，使得文本对抗样本的生成速度和攻击效果都有更进一步的提升。

另外一些提升生成样本质量的方法通过引入候选词池、WordNet 等外部知识库来对关键词进行近义替换[32,38],文献[37]在利用粒子群算法搜索近义词的同时也使用 HowNet 库作为支撑。但即使是近义单词也存在固定搭配、约定用法等上下文相关性带来的差异,进而导致生成的样本仍然可能存在语法、语义错误,并且攻击效果还可能受到词库规模的限制。因此,如何构建大规模、高质量的知识库将是该分支下一步的研究重点。

4.4　语句级对抗样本

基于深度学习的计算机视觉模型对于输入图像数据的形状(长、宽和通道深度)有着严格的限制,这使得在构造图像对抗样本时以改动现有图像的元素为主。文本数据的长度在上限内是可变的,因此可为原始文本增加新的干扰语句或对原始语句进行改写以生成语句级对抗样本。

对于增加干扰语句的方法,相关研究有 Jia 等[39]提出的两种针对机器问答的对抗样本生成方法,即 ADDSENT 和 ADDANY。ADDSENT 具体可分为 3 个阶段:1)对原问题中的名词、形容词进行反义词替换,并将命名实体和数字等 QA 中重要的文本对象替换词嵌入向量相近的对象;2)依据原答案的格式和实体类型产生假答案;3)将转换后的问题和答案合并为陈述句,并进行人工语法纠错,作为干扰语句与原文合并构成对抗样本。其最终使得 QA 系统的 F1 分数由 75%下降至 36%。ADDANY 使用贪心搜索方法寻找显著降低模型效果的修改词并生成扰动语句,同时去除了 ADDSENT 的语法正确性约束,使得模型性能进一步下降 7%。Minervini 等[40]尝试将对抗样本的生成转化为优化问题,这能够为文本蕴含等自然语言推理数据增加干扰背景知识,从而误导模型出错。该类方法攻击难度低、易操作,但是生成的对抗样本往往会引入新语句,导致扰动幅度过大,因此易被察觉,很难对短文本数据发起隐蔽攻击。

"复述(Paraphrases)"是指用与原句不同的词汇、句式重新写一个含义相同的句子的方法,相比增加干扰的方法更具隐蔽性,人类几乎无法区分生成的样本和原始样本。该分支的研究分歧点主要在于是采用自动化生成模型还是利用人机交互的方式来生成对抗样本。

关于自动化生成方法的主要研究有 Cheng 等[41]借鉴生成对抗网络的思想,提出基于 encoder-decoder 结构和平移损失的白盒方法 AdvGen,其中,编码器对输入语句进行复述以生成近似句作为对抗样本,解码器中的对抗性输入可用于训练防御攻击。实验证明,该方法能够对机器翻译系统进行攻击和加固。同样采用 encoder-decoder 结构的 SCPNs[42]接收"源语句-目标句法"对作为输入,通过编码器对原语句进行编码后使用回译(back-translation)方法生成近似语句,并将近似语句和目标句法输入解码器以生成对抗样本,成功实现了对文本分类和文本蕴含模型的攻击。Zhao 等[43]使用对抗性正则化自动编码器(Adversarially Regularized Autoencoders,ARAE)技术实现了文本离散编码到连续数据的转化,并在此基础上使用 WGAN[44]通过对抗学习的方法生成近似句,利用该方法得到的对抗样本被证明较为自然且符合人类语法规范。

尽管自动化方法具有效率高、成本低等特点,但是其主要的缺陷在于生成的样本质量不尽如人意,如 SCPNs[42]方法生成的结果会出现语句实体被破坏等异常情况,同时生成的目标主要是单句和短句,不能针对整个段落进行改写。Wallace 等[45]提出了一种"人类参与循环"的人机结合的方法,其不仅能够针对整个长段落实施复述,甚至解决了自动化方法无法解决的多跳推理难题,但付出的代价则是昂贵的众包成本和降低生成速度。

考虑到自动化方法和人机交互方法各有所长,一些研究尝试将二者结合。Ribeiro 等[46]进行了进一步的研究并提出了 SEA(Semantically Equivalent Adversaries)方法,为生成样本增加了约束条件 $SemEq(x,x')$,使其与原始文本保持语义等价,如式(17)和(18)所示,其中 τ 为阈值超参数。

$$SemEq(\boldsymbol{x},\boldsymbol{x}')=I(\min(1,\frac{p(\boldsymbol{x}'|\boldsymbol{x})}{p(\boldsymbol{x}|\boldsymbol{x})})\geqslant\tau) \tag{17}$$

$$SEA(\boldsymbol{x},\boldsymbol{x}')=I(SemEq(\boldsymbol{x},\boldsymbol{x}')\wedge f(\boldsymbol{x}')\neq f(\boldsymbol{x})) \tag{18}$$

同时他们基于 SEA 设计出通用的对抗样本生成规则 SEARs。该方法既可通过自动化的方式对句子进行同义改写,也可利用人机交互的方式生成更高质量的对抗样本,与文献[45]的研究具有共性。

4.5　多级对抗样本

一些经典的文本对抗方法并不拘泥于上述对字、词或句的分类,能够同时适用于生成不同级别的对抗文本或者尝试使用多个级别扰动相互配合的方法,代表性的方法有文献[47-50]中提出的方法。TEXTBUGGER[47]采用类似文献[30]的白盒方法来完成字和词级的扰动,并将对抗样本的搜索约束在上下文相关的词向量空间的 top-k 近邻中,以避免语义异常。HotFlip[48]扩展了文献[28]中提出的方法以生成词级对抗样本,并保留了生成字级样本的特性。Vijayaraghavan 等[49]提出 encoder-decoder 结构的 AEG 模型,通过采用深度强化学习的方式解决了离散数据生成的难题,并使用对抗奖励、语义相似奖励与词汇相似奖励之和作为训练奖励,以保证生成文本与原始文本的语义一致性,但其攻击效果受强化学习算法训练不稳定的影响而略低于传统方法,未来可考虑与其他方法结合以兼顾文本质量和攻击效果。此外,值得关注的是,Liang 等[50]提出的方法不仅能够生成字符级和词组级的对抗文本,还能完成黑盒和白盒两种攻击方式,具有较广泛的应用场景。

4.6　总结

本节所讨论的关键文本对抗样本技术的汇总如表 2 所列。各类面向 NLP 系统的对抗样本不仅具有良好的攻击效果,其中一些还具备攻击多类任务的能力。同时,白盒攻击方法虽然在很大程度上更具破坏力,但可被梯度掩蔽等措施防御,而黑盒方法能够让攻击者在知道少量或不知道模型细节信息的情况下轻松发动对抗攻击,从而进一步为检测和防御工作带来了挑战。

表 2 NLP 对抗样本生成方法汇总

Table 2 Summation of generation methods of NLP adversarial samples

级别	方法	任务	语言	已知信息	目标	数据集	指标	表现
char	Belinkov[25]	MT	en,de,cs	black-box	no	TED	BLEU	15↓
char	DeepWordBug[26]	TC	en	black-box	no	AG news	Acc	LSTM:68%↓ CNN:48%↓
char	WordHandling[27]	TC	zh	black-box	no	Hotel Reviews	Acc	35%↓
char	Ebrahimi[28]	MT	en,de,cs	black-box	both	TED	BLEU	21↓
char	VIPER[29]	Multi	en	black-box	no	G2P	模型效果下降 82%*	
word	Papernot[30]	TC	en	white-box	no	Movie Reviews	Asr	100%
word	TEXTFOOLER[32]	TC,TE	en	black-box	no	IMDB	Asr	99.7%
word	Samanta[33]	TC	en	white-box	no	IMDB	Acc	42%↓
word	iAdvT-Text[34]	TC	en	white-box	no	IMDB	Asr	94.34%
word	MHA[35]	TC	en	both	both	IMDB	Asr	Black:98.7% White:99.9%
word	Alzantot[36]	TC	en	black-box	yes	IMDB	Asr	97%
word	Zang[37]	TC	en	black-box	yes	IMDB	Asr	LSTM:100% BERT:98.7%
word	PWWS[38]	TC	en	white-box	no	IMDB	Acc	81.05%↓
sentence	Jia[39]	QA	en	black-box	no	SQuAD	F1	45%↓
sentence	Minervini[40]	TE	en	white-box	no	SNLI	*	*
sentence	AdvGen[41]	MT	Multi	white-box	no	NIST dataset	*	*
sentence	SCPNs[42]	TC,TE	en	black-box	no	SST	Acc	62.9%↓
sentence	GAN[43]	HT,MT,TE	en	black-box	no	SNLI	81%的受试者认为生成样本与原句相似	
sentence	Wallace[45]	QA	en	white-box	no	Quizbowl questions	Acc	21.7%↓
sentence	SEAs[46]	QA	en	black-box	no	VisualQA	Err	45.7%
multi	TEXTBUGGER[47]	TC	en	both	no	IMDB	Asr	100%
multi	HotFlip[48]	TC	en	white-box	yes	AG news	Asr	90%以上
multi	AEG[49]	TC	en	black-box	no	IMDB	Acc	71.45%↓
multi	Liang[50]	HT	en	both	both	*	受试者仅发现 40%的扰动	

注:表中只列举了部分具有代表性的任务和结果作为示例,* 表示原文未详细提供相关信息,↓表示下降数值,Err 表示错误率(Error Rate),HT 表示人类参与测试(Human Testing)

5 防御方法

相比对抗攻击方法,防御方法的研究仍处于起步阶段,并主要在数据、模型和推理这 3 个阶段进行防御和增强模型鲁棒性。

5.1 数据阶段

数据阶段可以通过数据预处理或为训练数据增加对抗样本等与模型无关的措施来增加模型鲁棒性,从而达到防御对抗样本的目的。一方面,上述对抗样本生成方法大多证明了利用生成对抗样本扩充训练数据集能够帮助模型在面对对抗攻击时变得更鲁棒,相关研究在文献[51]中被进一步引申为"对比集"理论,但这可能会增加模型的训练成本且通用性不强;另一方面,可以通过在数据流入 NLP 模型前进行预处理操作,对可能产生对抗攻击的元素进行修改或消除。文献[25]发现利用拼写纠正能够消除字符级扰动;Pruthi 等[52]在此基础上提出带回退的 ScRNN 模型,能够对与上下文含义冲突的疑似对抗样本进行修正,并提供了可灵活调整的灵敏度阈值。

然而,上述方法无法应对词或句级的对抗攻击。为了解决这一难题,Zhou 等[53]提出了识别扰动框架 DISP(discriminate perturbations)。如图 5 所示,该框架包含两个关键结构:1)扰动区分器,利用二分类问题计算文本中单词为扰动元素的可能性;2)词嵌入评估器,根据上下文信息和语料知识库的支持,使用 K 近邻法搜索合适的词嵌入进行扰动校正。其实现了对词组级别对抗样本的识别和防御。

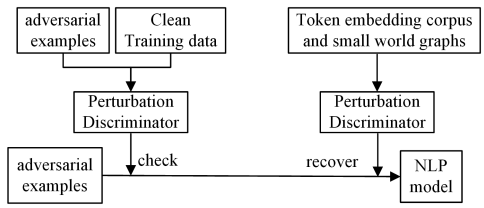

图 5 DISP 的防御框架

Fig. 5 Schematic of DISP defense framework

5.2 模型和推理阶段

模型阶段主要通过调整网络结构和修改训练方法等来提升模型的泛化能力和鲁棒性。从调整网络结构的角度来看,一些常用的可避免过拟合的措施(如 L2 正则化[54])或在输出时减少模型对用户反馈的梯度与结果置信度等关键信息经实验证明有助于减少对抗样本对模型的影响。蒸馏防御[55]利用训练好的原始模型(教师模型)的输出结果(软标签)代替数据集中的硬标签来训练一个结构完全相同的模型(学生模型),并通过在学生模型的 $softmax$ 层加入温度超参数 T 来增加输出结果的均匀性并平滑训练梯度,如式(19)所示,最终减少扰动对模型的干扰。

$$softmax(\boldsymbol{x},T)_i=\frac{e^{\frac{x^i}{T}}}{\sum_j e^{\frac{x^j}{T}}} \tag{19}$$

从修改训练方法的角度看,对抗训练是一种引入噪声的训练方式,可以对参数进行正则化以提升模型的鲁棒性和泛化能力。假设为输入加上导致损失值增大的扰动 $\boldsymbol{\varepsilon}$ 后,输出

分布和原标签 y 的分布一致,具体可通过修改损失函数来实现,如式(20)所示:

$$loss=-\log P(y|\boldsymbol{x}+\boldsymbol{\varepsilon}) \tag{20}$$

文献[56-58]分别提出了针对文本序列模型的经典对抗训练方法 FGM,PGD 和使用全局累积的关键词梯度信息的对抗训练方法 TextAT,这些方法在一定程度上受到对抗样本生成算法的启发。同时,Dinan 等[59]借鉴对抗训练的思想,提出一种人工参与的"构建模型-攻击模型-修复模型"的管道方法,实验结果证明其能够帮助 BERT 等模型抵御对抗样本的攻击。文献[41]将对抗样本和对抗训练相结合,使得对抗样本和真实样本经过编码器处理后的嵌入向量足够接近,并使二者经过解码后输出相同的机器翻译结果,以消除对抗样本的影响。

推理阶段的主流方法是借助集成学习的方式,将多种防御模型组合以输出更稳定的结果,但防御效果仍然有待进一步的验证[60]。

此外,评估 RNN[61],Transformer[62] 等模型鲁棒性的方法也可用于检测模型的安全性。但总体来说,上述方法仍只能抵挡部分对抗攻击,甚至会造成 NLP 模型的性能下降。如何探索更加高效、可靠的防御方法将始终是对抗样本领域研究的重难点。

6　挑战和前景展望

面向 NLP 的对抗样本技术的研究正处于起步阶段,因此仍有待完善和丰富。对于攻击方来说,存在以下难题。

(1)重理论而轻实战。当前 NLP 领域的对抗样本研究主要停留在理论层次,即用自己的攻击方法对抗自己训练的靶机模型。这些对攻击条件要求苛刻的白盒和黑盒攻击方法能否被实战应用于攻击现实世界中的 NLP 系统或检测其安全性仍有待研究。

(2)易被检测。不同于图像对抗样本的高隐蔽性特点,目前生成的文本对抗样本包含较多语法、语义错误,容易被人类或拼写系统识别和修复,而现有的对抗样本质量的衡量指标却无法体现这些扰动的显著性和易检测性。

(3)计算复杂度较高。大多数对抗样本生成方法的良好攻击效果是以计算复杂度为代价的。以基于词序评分机制的黑盒方法为例,攻击单个文本所需的模型推理次数与该文本的长度成正比,从而为攻击 LSTM 等计算速度较慢的神经网络模型带来了困难。

(4)跨语言攻击的迁移性难题。上述研究具有较强的语言相关性,不同自然语言之间的差异使得常用的英文字母替换、顺序改变等攻击策略很难被迁移至其他语言,从而限制了目前对抗样本生成方法的跨语言通用性。

对于防守方而言,面临的困境和挑战如下。

(1)数据集和工具匮乏。为训练集增加对抗样本和对抗训练仍然是目前增强模型鲁棒性的主流方法,但目前 NLP 领域缺乏大规模的、涵盖多种生成方法的对抗样本数据集,同时也没有类似图像领域中可用于生成对抗样本以进行对抗训练的 AdvBox[63] 等集成化工具。

(2)安全性和易用性的权衡。数据集扩充、对抗训练和蒸馏防御等虽然能够防御攻击,但增加了模型的训练代价并可能对模型性能产生负面影响。同样,隐藏梯度和置信度反馈的方法虽然避免了模型敏感信息泄露,但可能不能满足使用者的需求。此外,对抗样本检测方法也存在误报、漏报问题。如何权衡模型的安全性和易用性成为该领域的一大难题。

(3)没有永远的安全。受到神经网络自身固有脆弱性的影响,增加防御措施后的模型未必就是安全的。图像领域的一些研究证明,看似可靠的防御方法仍存在被攻破的风险[64]。

未来的工作除了聚焦于攻克上述难题,还将在通用性、隐蔽性和欺骗性等方面进一步拓展,具体可能的方向如下。

(1)中文对抗样本。目前的文本对抗样本研究主要围绕英文开展,而中文作为基于象形文字的语言,其语言结构、语法等特征与以英文为代表的表音文字迥然不同,针对中文的对抗样本的生成和防御方法也将别具一格。

(2)更多的通用文本对抗样本。针对文本的对抗样本生成方法在同语言内部大都具备可迁移性,能够对多种模型造成影响,这使得寻找更多能够攻击所有任务和所有 NLP 模型甚至能够对抗防御的特定语句作为通用对抗样本[65]成为可能,这些生成好的通用对抗性文本可被轻松分发和使用以对抗 NLP 系统。

(3)利用隐写术隐藏对抗样本。虽然文本对抗样本相比图像对抗样本更容易被人类识别,但隐写术为构造隐蔽文本对抗样本提供了途径,如将对抗文本设置为与文档、网页等载体背景色彩一致的白色字体或使用最小的字号,即可对其隐藏并轻松绕过人类视觉检测,最终对 NLP 系统发起攻击。

(4)面向人类的文本对抗样本。构造面向机器学习模型的文本对抗样本并非难事,在 GAN 技术的支持下,生成能够欺骗人类的文本对抗样本也不再是"天方夜谭"。深度伪造(Deepfake)技术[66]目前能够实现图像、视频和音频等信息载体的修改或生成并达到以假乱真的程度,该技术和 SeqGAN[67]等文本生成模型相结合,能够模仿目标人物的行文风格并生成难辨真假的文本,从而为人工智能安全带来了新挑战。

结束语　基于深度学习的 NLP 技术在诈骗短信识别、网络有害信息发现等生产、生活领域的应用愈发广泛,其本身的安全性问题也应引起大家的关注。本文对造成 NLP 模型脆弱性的原因进行了分析,并归纳总结了面向 NLP 的文本对抗样本的概念、特点、评估以及生成和防御方法。可以看到,现有研究能够针对英文轻松地生成字、词、句级别的扰动,但面向中文的对抗方法仍然很有限,一方面,这些攻击方法存在着未进行实战使用、易被检测和时间复杂度较高等不足;另一方面,相关防御技术仍然处于起步阶段,具备可迁移性甚至是通用性的对抗样本在融合隐写术后能发动更加隐蔽的攻击,进一步加重了模型安全性问题的负担。探索高效、可靠的防御方法对维护基于深度学习的 NLP 系统的安全性和良好发展将有着深远的影响。

参 考 文 献

[1]　HOCHREITER S,SCHMIDHUBER J. Long Short-Term Mem-

ory[J]. Neural computation, 1997, 9(8): 1735-1780.
[2] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient Estimation of Word Representations in Vector Space[J]. arXiv: 1301.3781, 2013.
[3] PENNINGTON J, SOCHER R, MANNING C. Glove: Global Vectors for Word Representation[C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014: 1532-1543.
[4] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [J]. arXiv: 1810.04805, 2018.
[5] YANG Z, DAI Z, YANG Y, et al. XLNet: Generalized Autoregressive Pretraining for Language Understanding [C] // Advances in Neural Information Processing Systems. 2019: 5754-5764.
[6] WANG W, WANG L, TANG B, et al. Towards a Robust Deep Neural Network in Text Domain A Survey [J]. arXiv: 1902.07285, 2019.
[7] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[J]. arXiv: 1312.6199, 2013.
[8] PAN W B, WANG X Y. Survey on Generating Adversarial Examples[J]. Journal of Software, 2020, 31(1): 67-81.
[9] ASHISH V. Attention is all you need[C]// Advances in Neural Information Processing Systems. 2017: 5998-6008.
[10] NIVEN T, KAO H Y. Probing Neural Network Comprehension of Natural Language Arguments[J]. arXiv: 1907.07355, 2019.
[11] KUSNER M, SUN Y, KOLKIN N, et al. From word embeddings to document distances[C] // International Conference on Machine Learning. 2015: 957-966.
[12] HUANG G, GUO C, KUSNER M J, et al. Supervised Word Mover's Distance[C] // Advances in Neural Information Processing Systems. 2016: 4862-4870.
[13] WU L. Word mover's embedding: From word2vec to document embedding[J]. arXiv: 1811.01713, 2018.
[14] DONG Y, FU Q A, YANG X, et al. Benchmarking Adversarial Robustness[J]. arXiv: 1912.11852, 2019.
[15] MICHEL P, LI X, NEUBIG G, et al. On Evaluation of Adversarial Perturbations for Sequence-to-Sequence Models [J]. arXiv: 1903.06620, 2019.
[16] GIANNA M D C, ANTONIO G, FRANCESCO R, et al. Ranking a stream of news[C]// Proceedings of the 14th International Conference on World Wide Web. 2005: 97-106.
[17] RICHARD S. 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.
[18] CETTOLO M, GIRARDI C, FEDERICO M. Wit3: Web inventory of transcribed and translated talks[C]// Conference of European Association for Machine Translation. 2012: 261-268.
[19] RAJPURKAR P, ZHANG J, LOPYREV K, et al. SQuAD: 100 000+ Questions for Machine Comprehension of Text[J]. arXiv: 1606.05250, 2016.
[20] RAJPURKAR P, JIA R, LIANG P. Know What You Don't Know: Unanswerable Questions for SQuAD[J]. arXiv: 1806.03822, 2018.
[21] GOYAL Y, KHOT T, SUMMERS-STAY D, et al. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017: 6904-6913.
[22] BOWMAN S R, ANGELI G, POTTS C, et al. A large annotated corpus for learning natural language inference[J]. arXiv: 1508.05326, 2015.
[23] WILLIAMS A, NANGIA N, BOWMAN S R. A broad-coverage challenge corpus for sentence understanding through inference [J]. arXiv: 1704.05426, 2017.
[24] ERIK F, SANG T K, DE MEULDER F D. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition[J]. arXiv: 0306050, 2003.
[25] BELINKOV Y, BISK Y. Synthetic and natural noise both break neural machine translation[J]. arXiv: 1711.02173, 2017.
[26] 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). IEEE, 2018: 50-56.
[27] WANG W Q, WANG R. Adversarial Examples Generation Approach for Tendency Classification on Chinese Texts[J]. Journal of Software, 2019, 30(8): 2415-2427.
[28] EBRAHIMI J, LOWD D, DOU D. On adversarial examples for character-level neural machine translation [J]. arXiv: 1806.09030, 2018.
[29] EGER S, ŞAHIN G G, RÜCKLÈ A, et al. Text processing like humans do: Visually attacking and shielding NLP systems[J]. arXiv: 1903.11508, 2019.
[30] PAPERNOT N, MCDANIEL P, SWAMI A, et al. Crafting adversarial input sequences for recurrent neural networks[C]// MILCOM 2016-2016 IEEE Military Communications Conference. IEEE, 2016: 49-54.
[31] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[J]. arXiv: 1412.6572, 2014.
[32] 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[J]. AAAI2020, arXiv: 1907.11932, 2019.
[33] SAMANTA S, MEHTA S. Towards crafting text adversarial samples[J]. arXiv: 1707.02812, 2017.
[34] SATO M, SUZUKI J, SHINDO H, et al. Interpretable adversarial perturbation in input embedding space for text[J]. arXiv: 1805.02917, 2018.
[35] ZHANG H, ZHOU H, MIAO N, et al. Generating Fluent Adversarial Examples for Natural Languages[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 5564-5569.
[36] ALZANTOT M, SHARMA Y, ELGOHARY A, et al. Generating natural language adversarial examples [J]. arXiv: 1804.07998, 2018.
[37] ZANG Y, YANG C, QI F, et al. Textual Adversarial Attack as Combinatorial Optimization[J]. arXiv: 1910.12196, 2019.
[38] 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 Associa-

tion for Computational Linguistics. 2019:1085-1097.
[39] JIA R,LIANG P. Adversarial examples for evaluating reading comprehension systems[J]. arXiv:1707.07328,2017.
[40] MINERVINI P,RIEDEL S. Adversarially regularising neural nli models to integrate logical background knowledge[J]. arXiv:1808.08609,2018.
[41] CHENG Y,JIANG L,MACHEREY W. Robust neural machine translation with doubly adversarial inputs[J]. arXiv:1906.02443,2019.
[42] IYYER M,WIETING J,GIMPEL K,et al. Adversarial example generation with syntactically controlled paraphrase networks[J]. arXiv:1804.06059,2018.
[43] ZHAO Z,DUA D,SINGH S. Generating natural adversarial examples[J]. arXiv:1710.11342,2017.
[44] ARJOVSKY M,CHINTALA S,BOTTOU L. Wasserstein gan[J]. arXiv:1701.07875,2017.
[45] WALLACE E,RODRIGUEZ P,FENG S,et al. Trick me if you can:Human-in-the-loop generation of adversarial examples for question answering[J]. Transactions of the Association for Computational Linguistics,2019,7(2019):387-401.
[46] RIBEIRO M T,SINGH S,GUESTRIN C. Semantically equivalent adversarial rules for debugging nlp models[C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. 2018:856-865.
[47] LI J,JI S,DU T,et al. Textbugger:Generating adversarial text against real-world applications[J]. arXiv:1812.05271,2018.
[48] EBRAHIMI J,RAO A,LOWD D,et al. Hotflip:White-box adversarial examples for text classification[J]. arXiv:1712.06751,2017.
[49] VIJAYARAGHAVAN P,ROY D. Generating Black-Box Adversarial Examples for Text Classifiers Using a Deep Reinforced Model[J]. arXiv:1909.07873,2019.
[50] LIANG B,LI H,SU M,et al. Deep text classification can be fooled[J]. arXiv:1704.08006,2017.
[51] GARDNER M,ARTZI Y,BASMOVA V,et al. Evaluating nlp models via contrast sets[J]. arXiv:2004.02709,2020.
[52] PRUTHI D,DHINGRA B,LIPTON Z C. Combating adversarial misspellings with robust word recognition[J]. arXiv:1905.11268,2019.
[53] ZHOU Y,JIANG J Y,CHANG K W,et al. Learning to discriminate perturbations for blocking adversarial attacks in text classification[J]. arXiv:1909.03084,2019.
[54] TANAY T,GRIFFIN L D. A New Angle on L2 Regularization[J]. arXiv:1806.11186,2018.
[55] PAPERNOT N,MCDANIEL P,WU X,et al. Distillation as a defense to adversarial perturbations against deep neural networks[C]// 2016 IEEE Symposium on Security and Privacy (SP). IEEE,2016:582-597.
[56] MIYATO T,DAI A M,GOODFELLOW I. Adversarial training methods for semi-supervised text classification[J]. arXiv:1605.07725,2016.
[57] MADRY A,MAKELOV A,SCHMIDT L,et al. Towards deep learning models resistant to adversarial attacks[J]. arXiv:1706.06083,2017.
[58] LI L,QIU X. TextAT:Adversarial Training for Natural Language Understanding with Token-Level Perturbation[J]. arXiv:2004.14543,2020.
[59] DINAN E,HUMEAU S,CHINTAGUNTA B,et al. Build it break it fix it for dialogue safety:Robustness from adversarial human attack[J]. arXiv:1908.06083,2019.
[60] HE W,WEI J,CHEN X,et al. Adversarial example defense:Ensembles of weak defenses are not strong[C]// 11th USENIX Workshop on Offensive Technologies (WOOT 17). 2017.
[61] KO C Y,LYU Z,WENG T W,et al. POPQORN:Quantifying robustness of recurrent neural networks[J]. arXiv:1905.07387,2019.
[62] SHI Z,ZHANG H,CHANG K W,et al. Robustness verification for transformers[J]. arXiv:2002.06622,2020.
[63] GOODMAN D,XIN H,YANG W,et al. Advbox:a toolbox to generate adversarial examples that fool neural networks[J]. arXiv:2001.05574,2020.
[64] ATHALYE A,CARLINI N,WAGNER D. Obfuscated gradients give a false sense of security:Circumventing defenses to adversarial examples[J]. arXiv:1802.00420,2018.
[65] WALLACE E,FENG S,KANDPAL N,et al. Universal adversarial triggers for nlp[J]. arXiv:1908.07125,2019.
[66] LIANG R G,LYU P Z,et al. A Survey of Audiovisual Deepfake Detection Techniques[J]. Journal of Cyber Security,2020,5(2):1-17.
[67] YU L,ZHANG W,et al. Seqgan:Sequence generative adversarial nets with policy gradient[C]// Thirty-First AAAI Conference on Artificial Intelligence. 2017.

TONG Xin, born in 1995, postgraduate, is a member of China Computer Federation. His main research interests include adversarial examples and natural language processing.

WANG Bin-jun, born in 1962, professor, Ph. D supervisor, is a member of China Computer Federation. His main research interests include natural language processing and information security.