文章编号：1003-0077(2021)08-0001-15

文本对抗样本攻击与防御技术综述

杜小虎[1]，吴宏明[2]，易子博[1]，李莎莎[1]，马 俊[1]，余 杰[1]

（1. 国防科技大学 计算机学院，湖南 长沙 410073；
2. 中央军委装备发展部 装备项目管理中心，北京 100034）

摘 要：对抗样本攻击与防御是最近几年兴起的一个研究热点，攻击者通过微小的修改生成对抗样本来使深度神经网络预测出错。生成的对抗样本可以揭示神经网络的脆弱性，并可以修复这些脆弱的神经网络以提高模型的安全性和鲁棒性。对抗样本的攻击对象可以分为图像和文本两种，大部分研究方法和成果都针对图像领域，由于文本与图像本质上的不同，在攻击和防御方法上存在很多差异。该文对目前主流的文本对抗样本攻击与防御方法做出了较为详尽的介绍，同时说明了数据集、主流攻击的目标神经网络，并比较了不同攻击方法的区别。最后总结文本对抗样本领域面临的挑战，并对未来的研究进行展望。

关键词：自然语言处理；对抗样本；深度神经网络

中图分类号：TP391 **文献标识码**：A

Adversarial Text Attack and Defense: A Review

DU Xiaohu[1], WU Hongming[2], YI Zibo[1], LI Shasha[1], MA Jun[1], YU Jie[1]

(1. School of Computer Science, National University of Defense Technology, Changsha, Hunan 410073, China;
2. Equipment Project Management Center of Equipment Development Department,
Central Military Commission, Beijing 100034, China)

Abstract: Adversarial attack and defense is a popular research issue in recent years. Attackers use small modifications to generate adversarial examples to cause prediction errors from the deep neural network. The generated adversarial examples can reveal the vulnerability of the neural network, which can be repaired to improve the security and robustness of the model. This paper gives a more detailed and comprehensive introduction to the current mainstream adversarial text example attack and defense methods, the data set together with the target neural network of the mainstream attack. We also compare the differences between different attack methods in this paper. Finally, the challenges of the adversarial text examples and the prospect of future research are summarized.

Keywords: natural language processing; adversarial example; deep neural network

1 介绍

1.1 对抗样本

深度神经网络在解决各种现实任务中有着广泛的应用并取得了不错的效果，如计算机视觉[1-2]、图像分类[3]和自然语言处理[4-5]等领域。但是，近年来，人们开始关注神经网络的安全性与鲁棒性问题[6]，神经网络容易受到对抗样本的攻击，例如，指纹识别[7]，攻击者可以利用对抗样本来做伪装，非法破解指纹识别器。在垃圾邮件检测[8]中，攻击者也可以通过伪装逃避检测。在 NLP 任务中，神经网络可能受到经过精心修改的文本欺骗，这种修改后的文本与原始文本有着相当高的相似性，鲁棒的神经网络模型会对相似的文本做出与原文本一致的预测，但神经网络会对这种修改后的文本做出文本预测错误，这种经过微小修改后的文本称为对抗样本，使用这些对抗样本使神经网络发生错误的过程叫做对抗攻击。

收稿日期：2020-07-09 **定稿日期**：2020-09-18

基金项目：国家重点研究与发展计划(2018YFB1004502)

1.2 相关术语

鲁棒性(robustness)：鲁棒性[9]反映了自然语言处理模型对不同输入样本的适应程度，一个鲁棒性高的模型更加不易受到对抗样本的攻击。模型在意义相似或相近的输入文本上能够表现出一致的性能。

词向量(word embedding)：也叫词嵌入，由于深度神经网络的输入只能是连续的向量，所以其不能处理实际的文本，词向量即是文本在词空间中映射出的由实数组成的向量。

扰动(perturbation)：对原始文本做出的细微修改称为扰动，如对文本中单词的修改、删除和增加等操作。

1.3 NLP任务介绍

自然语言处理(NLP)是人工智能领域下的一个重点和热点研究方向。近年来，深度神经网络(DNNs)被应用于大量的NLP任务中，本文重点介绍一些用到深度神经网络且涉及对抗样本攻击与防御的NLP任务。

分类任务：包括文本分类[10]、情感分析[11]等。任务的目的是根据文本的内容给出预测的标签，这些分类任务可以是二分类，也可以是多分类。如情感分类中，将电影评论分为积极的与消极的；新闻分类中把不同的新闻分类到不同的新闻类别，如娱乐新闻、军事新闻等。

文本蕴含[12]：本任务是针对两段文本的关系的判断，其中一个为前提，另一个为假设。如果文本P推出文本Q，即为P蕴含Q，两者为蕴含关系，否则，两者为矛盾关系。

机器翻译[13]：把一段文本从一种语言翻译为另一种语言，如常见的中英文之间互相翻译。

问答系统：从特定的上下文文本中提取信息构造模型来回答人类用自然语言提出的问题。

1.4 文本对抗样本与图像的区别

对抗样本起源于图像，图像的对抗样本有着肉眼完全不可见的效果，如只修改图像的一个像素，这种扰动人类不易察觉，但神经网络能把修改后的图像判断为错误图像，这就是对抗样本最原始的目的。由于图像是连续的，可以将很微小的扰动通过搜索或者其他方法引入图像中，而这种扰动对人类是不可见的。但是文本是一个离散的序列，任何对文本的修改都可能引起人们的注意，例如，添加字符或替换单词[14]。同时这些改变可能改变文本原有的语义，例如，在句子中加入“not”，类似的否定词会改变句子语义，如果在情感分类任务中也会改变句子情感倾向。Liang等人[15]的实验表明了将在图像上表现较好的FGSM(fast gradient sign method)直接应用在文本攻击上会产生混乱不可读的对抗样本，虽然这种对抗样本成功地使分类器判别错误，但与原始文本差异明显，并不可取。文本的对抗样本生成有两个思路，一是跟图像一样做尽量小的微小修改，让人们尽可能地发现不了这种修改，类似于人们自己可能发生的错误，如单词拼写错误、键盘误触使单词出错。这种主要是字符上的修改。另一个思路是不像图像那样产生人类完全不可见的修改。而是产生人类判断正确却会使神经网络预测错误的样本。这就需要考虑两个问题：一是修改部分在语法和语义上与原文本需要有很大的相似性；二是修改的比例不能过高，修改过多会使文本失去原有的语义。这种情况主要体现在单词级别的修改。

1.5 词空间中文本相似度量方法

生成的对抗样本不仅必须欺骗目标模型，而且还必须使人们检测不到扰动。一个好的对抗样本应传达与原始文本相同的语义，因此需要一些度量标准来量化相似性。本文介绍三个指标：欧氏距离、余弦相似度和词移距离[16]。

欧氏距离(euclidean distance)：在文本中，欧几里得距离即是计算两个词向量之间的线性距离。

余弦相似度(cosine similarity)：余弦相似度通过计算两个向量之间的角度的余弦值来表示两个单词的语义相似度。余弦距离更关心两个向量之间的方向差异。两个向量之间的角度越小(余弦值越大)，相似度越大。

词移距离(word movers distance)：WMD[17]主要反映文档之间的距离，因此不适合查找相似的单词。其语义表示可以是基于Word2Vec或其他方式获得的词向量。该算法将文档距离构造为两个文档中单词的语义距离的组合。例如，从对应于两个文档中任意两个单词的单词向量中获得欧氏距离，然后从权重和总和中获得。两个文本A和B之间的WMD距离如式(1)所示。

$$\mathrm{WMD}(A,B)=\sum_{i,j}^{n}T_{ij}\cdot D(\vec{i},\vec{j}) \tag{1}$$

其中，$D(\vec{i},\vec{j})$是对应于两个单词i和j的词向

量的欧氏距离。该指标使用词袋模型(the bag of words)获取文本中单词的单词频率来作为文本中单词的权重，然后问题就变成了如何将文档 A 的所有单词单元映射到相应的单词 B 文档的词单元成本最低，最终得到权重矩阵 $\boldsymbol{T}$。WMD 算法是 EMD[18]算法的特例。

1.6　常见数据集

IMDB[19]：实验数据集包含 50 000 条 IMDB 电影评论，其中，训练集 25 000 条，测试集 25 000 条，该数据集专门用于情感分析。评论的结果是二分类的。标签为 pos(积极)和 neg(消极)。在训练集和测试集中两种标签各占一半。

AG's News[20]：从新闻文章中提取的带有标题和描述信息的数据集。它包含 4 种新闻类别，每种新闻类别包含 30 000 个训练样本和 1 900 个测试样本。

Yahoo! Answers[20]：是一个具有 10 种类别的主题分类数据集，其中包含 44 83 032 个问题和相应的答案。在十种类别中每个类别具有 140 000 个训练样本和 5 000 个测试样本。

SNLI[21]：是文本蕴含任务的数据集，包含 57 万对人工标记为蕴含、矛盾或中性的人类书写的英文句子对。

DBpedia：文本分类数据集，包含 560 000 个训练样本和 70 000 个测试样本，这些样本来自 14 个类别，例如公司、建筑和电影等。

SST：是斯坦福大学发布的一个情感分析数据集，主要针对电影评论来做情感分类。有两个版本可以用。一个是包含五个标签的 SST-1，一个是二分类的 SST-2。SST-1 共有 11 855 条样本，其中 8 544 条训练样本，1 101 条验证样本，2 210 条测试样本。SST-2 有 6 920 条训练样本，872 条验证样本，1 821 条测试样本。

1.7　本文贡献

本文对文本对抗样本攻击与防御方法的前沿技术进行了调研，分别从白盒和黑盒攻击两个方面回顾了文本对抗攻击的发展历史，同时对现有的最新防御技术做了介绍。最后，我们总结了文本对抗样本领域如今面临的挑战，并做了前景展望。与现有的文本对抗样本综述[22-23]相比，本文的贡献如下：

(1) 现有的文本对抗样本综述文章对相关方法的具体公式描述较少，且没有统一的描述形式。本文统一了攻击和防御领域相关论文的所有符号和公式表示形式，特别是在黑盒攻击部分对于单词重要性分数计算部分做了公式统一编辑，使读者能够更好地比较各个方法的具体细节差异。

(2) 本文相较于现有的综述文章补充了最新的前沿技术。

(3) 本文是首次针对中文的文本对抗样本攻击和防御研究进行综述，分析了中文与英文的不同，中文的对抗样本研究也是一个重要的研究方向。

1.8　本文符号说明

由于不同的攻击方法中符号描述不尽相同，本文采用统一的符号来表述对抗样本中涉及的各项数据。表 1 展示了本文中所用到的符号及其描述。

表 1　符号描述表

符号	描述
x	原始输入文本
x'	原始文本对应的对抗样本
w_i	文本中第 i 个单词
δ	对原始文本的扰动
Y	原始文本的标签
$F_Y(x)$	输入的文本在模型 F 中对应 Y 标签的输出分数
$F(x)$	输入的文本在模型 F 中输出的标签，如 $F(x)=Y$ 说明 x 对应的标签为 Y
$x^{\|wi}$	文本删除 w_i 后的新文本

注：$x'=x+\delta$，$x=\{w_1,w_2,\cdots,w_n\}$。

2　文本对抗样本攻击

2014 年，Szegedy 等人[6]发现用于图像分类的深度神经网络可以被添加过微小像素扰动的图像所欺骗。实验表明，图像分类器有很高的误分类率，但人类没有检测到图像的这种变化。2017 年，Jia 等人[24]率先考虑在基于文本的深度神经网络上生成对抗样本。从那时起，人们开始关注文本的对抗样本。

2.1　文本对抗样本分类

对抗样本攻击根据是否了解模型参数可以分为两种，即白盒攻击和黑盒攻击。前者可以获得模型的参数和结构，而后者缺乏这些信息。黑盒攻击更

具挑战性,因为它们通常需要设计有效的搜索算法对模型进行大量查询[25]从而获得单词重要性分数,以此来决定单词的替换顺序。大多数白盒攻击的方法都是基于梯度在使模型损失函数最大化的方向上做扰动,当攻击者不了解模型参数和结构时,只能采用黑盒攻击。黑盒攻击通过采用不断的查询和观察目标模型的输出来产生最优的扰动。

根据扰动的位置不同可以分为字符级攻击和单词级攻击。字符级攻击即是扰动单词中的单个字符,可以是删除、增加、修改或者两个字符交换。单词级攻击是对文本中的单个单词做扰动,同样也有增加、删除、修改的操作,但单词级攻击一般不采用删除的方式。删除文本的单词有比较大的风险出现语法错误和语义不通顺,大多数的单词级攻击为增加和修改,修改以同义词替换为主,增加的方式有在文本开始或者结尾增加一段无关的文字,这样原有的文本不会受到任何破坏,但分类器会预测出错。表 2 展示了近几年常见的文本对抗样本攻击方法。

表 2 常见的文本对抗样本攻击方法对比

攻击方法	攻击条件	NLP 任务
HotFlip[26]	白盒字符级	文本分类
Ebrahimi[27]	白盒字符级	机器翻译
Liang[15]	白盒字符级	文本分类
Tsai[28]	白盒单词级	情感分类
human-in-loop[29]	白盒单词级	问答系统
DeepWordBug[30]	黑盒字符级	文本分类
Heigold[31]	黑盒字符级	机器翻译
VIPER[32]	黑盒字符级	文本分类
Jia[24]	黑盒单词级	阅读理解
WS[33]	黑盒单词级	情感分类
Greedy[28]	黑盒单词级	情感分类
Wee[34]	黑盒单词级	问答系统
PWWS[35]	黑盒单词级	情感分类
TEXTFOOLER[36]	黑盒单词级	文本分类
GA[37]	黑盒单词级	情感分类
IGA[38]	黑盒单词级	情感分类
PAWS[39]	黑盒单词级	释义对
PSO[40]	黑盒单词级	情感分类

2.2 文本对抗样本白盒攻击方法

2.2.1 字符级白盒攻击

在 2017 年,Ebrahimi 等人[26]提出了一种称为 HotFlip 的基于梯度的白盒攻击方法来生成对抗样本。该方法基于 one-hot 输入向量的梯度对字符做修改,包括替换、删除和增加字符。通过评估哪个字符修改的损失最大,并利用束搜索来寻找最优的修改。这种攻击针对 CharCNN-LSTM[41]模型在 AG's News 数据集上的表现优于贪心搜索攻击算法。表 3[26]展示了 HotFlip 的攻击效果,仅仅将 mood 单词中的 d 字符替换为 P 就使得模型将新闻的分类由 57%置信度的 World 误分类为 95%置信度的 Sci/Tech 类别。

表 3 原始文本和 HotFlip 生成的对抗样本

South Africa's historic Soweto township marks its 100th birthday on Tuesday in a mood of optimism. 57% **World**
South Africa's historic Soweto township marks its 100th birthday on Tuesday in a moo**P** of optimism. 95% **Sci/Tech**

第二年,Ebrahimi 等人扩展了 HotFlip 方法,改进了束搜索,提出了 one-shot 攻击[27],这种攻击并行操作所有单词,不需要全局对梯度排序,减少了 HotFlip 的训练时间,常规训练时间减少至 1/3。攻击对象由文本分类器变成了机器翻译(NMT)模型,使用 TED 平行语料库作为数据集。实验结果表明,白盒攻击的效果优于 Belinkov 等人[42]的黑盒攻击方法。

2018 年,Liang 等人[15]针对文本分类提出了一种字符级白盒攻击方法。该方法先通过梯度确定原始文本中对分类贡献较大的词组,将其称为热样本词组 HSPs(hot sample phrases),然后对这些 HSPs 词组做修改,修改过程是从相关的语料库中获取常见的拼写错误替换 HSP,如将 film 替换为 flim,或者用外观类似的字符替换,如将小写字母 l 替换为数字 1。这种白盒攻击的目标模型为 DNN 模型,使用数据集为 DBpedia。实验结果表明,该方法可以成功地使字符级神经网络分类器出错并将原始文本扰动到任何期望的分类类别,即目标攻击。同时这种扰动很难被感知。表 4[15]展示了该方法所生成的对抗样本,仅仅修改单词 film 就使得模型将文本分类的类别由 99.6%置信度的电影类别误分类为 99%置信度的公司类别。

表 4　拼写错误生成的对抗样本

Maisie is a comedy **fli**m property MGM originally purchased for Jean Harlow but before a shooting script could be completed Harlow died in 1937. (film)

2.2.2　单词级白盒攻击

Behjati 等人[43]提出了一种基于梯度的通用性对抗样本攻击，即生成可以添加到任何输入中的单词序列。实验表明，文本分类器很容易受到此类攻击的影响，在每个输入序列开头插入一个对抗词可使模型准确性由 93%下降到 50%。

Tsai 等人[28]指出贪心搜索攻击方法不能保证产生最优的结果，而且耗时太多，因为算法在每一个迭代中需要搜索候选词。与此同时，考虑到贪心搜索算法的本质，算法可能在较早位置替换对最终目标没有太大贡献的次优词。另一个限制是，被替换的词往往在句子的一个邻近区域，特别是在句子前面一部分。这大大降低了句子的可读性，甚至破坏了原文的语义。Tsai 等人[28]提出了一种“全局搜索”算法，通过计算梯度扰动幅度来获得候选词，然后在扰动较大的位置替换这些词。扰动越大，分类器对词的变化越敏感。结果表明，全局搜索的结果比贪婪攻击产生的对抗样本更好，攻击成功率更高。

在问答系统上，Eric 等人[29]提出通过 human-in-loop 的方式生成对抗样本，与传统攻击方式利用单词替换生成对抗样本并人工评估样本有效性不同，human-in-loop 加入了人工修改，使用人机协作方式生成更加多样的对抗样本。在攻击过程中，该方法基于梯度计算单词重要性分数，拥有更高的计算效率。通过用户界面做人机交互，在用户界面中，作者写出问题，模型会预测五个结果并解释这些预测的原因，如果最正确的预测是正确答案，界面会指出该问题在模型中的哪个位置正确，目的就是让模型出错或者延迟找出正确答案的位置，该界面会随着作者的修改不断更新。human-in-loop 首次将对抗攻击过程可视化，同时揭示了问答系统的局限性。

白盒攻击的相关工作已经证明该方法在基于文本的攻击中是非常有效的，但仍存在一些问题需要解决，例如，在单词相似度、语法正确性和语义相似度等方面存在一些不足，且应用场景有局限性。

2.3　文本对抗样本黑盒攻击方法

2.3.1　字符级黑盒攻击

Gao 等人[30]于 2018 年提出了一种称为 DeepWordBug 的字符级黑盒攻击方法。该方法使用一种新的评分策略来识别关键字符并排序。使用简单的字符替换排名最高的单词，以最小化扰动的编辑距离，并改变原始的分类。该方法在文本分类、情绪分析和垃圾邮件检测等任务中取得了良好的效果，并降低了目前最先进的深度学习模型的预测精度。

Heigold 等人[31]在机器翻译和形态标记上提出了黑盒攻击方法，该方法提出了三种扰动方式：字符交换、字符翻转、单词扰动。字符交换是在单词中随机交换两个相邻字符，字符翻转是将一个字符随机替换为另外一个字符，单词扰动为随机扰乱除第一个和最后一个字符外的其他字符。在形态标记任务上，当字符翻转和字符交换比例设置为 10%时，模型的准确性由 95%下降到了 80%左右，下降非常明显。在机器翻译上，采用单词扰动的方式，评估结果的 BLEU 分数由 30 下降为 5，也是大幅地下降。

2019 年，Eger 等人[32]提出了一种称为 VIPER 的字符级白盒攻击方法。它在视觉空间中寻找一个与原始文本中的字符最相似的字符并将其替换，实验结果表明，受到 VIPER 攻击的 SOTA 模型性能下降达 82%，但人们只受到轻微甚至是感受不到的扰动。与 HotFlip 产生的对抗样本容易造成不可读的情况不同，VIPER 方法理想情况下是可读的，该方法通过概率 p 和词空间 CES 来决定替换字符，对输入文本的每个字符做替换，如果发生替换，则是选择词空间中 20 个最邻近字符的一个。替换后的字符 w_i' 可表示为两个参数，如式(2)所示。

$$w'_i = \mathrm{VIPER}(p, \mathrm{CES}) \tag{2}$$

其中，选择字符 w_i' 的 20 个邻近词概率 p 与它们到 w_i' 的距离成正比，CES 可以是任何可用于识别字符邻近词的词空间。表 5[32]展示了 VIPER 的攻击效果，全文替换的对抗样本看上去差异明显，而少量字符的替换则完全不影响阅读，在 Facebook 和 Twitter 的有毒评论检测模型中就可能面临这样的对抗样本攻击，用户以这种相似的字符做伪装而逃避模型的检测，但是用户仍然表达了其观点，其他用户也可以完全看出文本原来的意思。

表 5　原始文本和 VIPER 生成的对抗样本

Original Text:	he is also a faggot .
Adversarial Text:	ĥe is also a faggot .
Adversarial Text:	he is alṣo a fagot .
Original Text:	Mr. Coffee is a professor at Columbia Law School.
Adversarial Text:	Ḿř. Ĉôf̈f̂ê îŝ â ṕřôfêŝsoř âẗ Ĉôl̨ŭḿb̂îâ Ĺâẃ Ŝĉĥôôl̨.

2.3.2 单词级黑盒攻击

单词级攻击的优点是能够很大程度地保持语义,且不会像字符级攻击那样产生不存在的单词。Jia等人[24]于2017年提出了一种针对阅读理解系统的黑盒攻击。作者提出的攻击方法是在段落末尾添加一些分散注意力但毫无意义的句子。这些分散注意力的句子不会改变段落和答案的语义,但它们会使神经网络模型出错。分散注意力的句子可以是一个精心生成的真实句子,也可以是一个使用20个随机的常见单词的任意单词序列。最后,对神经网络进行迭代查询,直到输出发生变化,即认为攻击成功。

2018年,Samanta等人[33]提出了一种叫作词显著性(word saliency,WS)的单词级黑盒攻击,通过对单词的删除、替换和增加等操作生成对抗样本。该方法先计算每个单词对分类结果的贡献程度,并把它们按照从大到小排序,如果某个单词贡献大且是副词,则删除这个词,在剩余的单词中找出每个单词的候选词,在候选词中选择对模型正确分类贡献程度最小的做替换。在替换时,如果被替换的单词是形容词且候选词是副词,则将候选词加到被替换单词后面,否则用候选词直接替换原词。单词 w_i 贡献率的计算,如式(3)所示。

$$C_F(w_i,Y)=\begin{cases}F_Y(x)-F_Y(x^{|w_i}),\\ \quad if\ F(x)=F(x^{|w_i})=Y\\ F_Y(x)+F_Y(x^{|w_i}),\\ \quad if\ F(x)=Y \text{ and } F(x^{|w_i})\neq Y\end{cases} \tag{3}$$

其中,$F_Y(x)$是文本 x 在分类器 F 中属于 Y 标签的概率,$x^{|w_i}$ 是文本 x 去除目标词 w_i 后的新文本。大量的查询分类是一个非常耗时的过程。在情感分类数据集IMDB中,生成的对抗样本与原始文本在Spacy工具的测试下有90%以上的相似度,实验结果还表明,更低的文本相似度会带来更多的有效对抗样本数量,这说明文本的相似性跟攻击成功率成反比。

Tsai等人[28]提出一种贪心搜索攻击方法(greedy search attack),在每次迭代中,算法都会计算出每个词在词空间中的 k 个近邻。然后从 k 个邻居中选出对预测结果影响最大的一个。虽然词向量有助于找到在相似上下文中使用的单词,但它不能保证替换后的词性保持不变。因此,该方法还检查了原词的词性,并确保所选的候选词具有与原词相同的词性。贪心搜索攻击中有三个因素会影响结果:一是 k 的大小,k 太小时,可能没有足够的候选词来形成一个对抗样本;二是表示候选单词与原单词在词空间中允许的最大距离 d。当 d 很大时,改变后的句子的语义一般会离原句更远。当 d 太小时,成功攻击的机会就会降低;三是表示句子中允许替换的百分比阈值 r。当 r 太小时,生成成功的对抗样本机会就会降低。当 r 太大时,会有太多的单词被替换,所以产生的样本与原句差别很大。

Wee等人[34]针对问答系统(QA)提出了黑盒攻击方法,并将生成的对抗样本作为数据扩充方法来重新训练模型,以避免在未知的测试数据上性能下降,同时提高模型的泛化能力和可靠性。这种攻击提出了一种新的训练神经网络方法,该方法将原问题和候选单词或短语作为输入来生成对抗样本,训练数据集包括了原问题、候选词/短语和目标问题,其中候选词/短语是目标问题的一部分。这种训练可以为给定的问题生成多个对抗样本。在人工评估方面采用来自AMT(Amazon mechanical turk)的人类注释器(human annotators)评估了对抗样本的等效性和流畅性。在3 000个对抗样本上认为对抗样本与原始问题含义一致的占78%,认为生成的对抗样本是流利的英文占78.6%。表6[34]展示了原始问题和对抗样本,可以看到将短语associated with替换为related to就引起了预测的变化。

表6 问答系统的对抗样本

Context: … According to the Second law of thermodynamics, nonconservative forces necessarily result in energy transformations within closed systems from ordered to more random conditions as entropy increases.
Original Question: What is the law of thermodynamics associated with closed system heat exchange? **Prediction**: Second law of thermodynamics
Paraphrased Question: What is the law of thermodynamics related to closed system heat exchange? **Prediction**: nonconservative forces

Ren等人[35]于2019年提出了概率加权词显著性PWWS(probability weighted word saliency)的黑盒攻击方法,PWWS方法在WS方法的基础上有所改进。先按照WS方法计算出文本 x 中的每个单词 w_i 的词显著性向量 $S(x)$,在确定替换单词的优先级时,考虑替换后的分类概率的变化程度和每个单词的显著性。以 $x_i^*=(w_1w_2\cdots w_i^*\cdots w_n)$ 表示用 w_i^* 替换 w_i 的文本,用 $\Delta P_i^*=F_Y(x)-F_Y(x_i^*)$ 表示 w_i 替换的显著性。最后通过如下函数定义

w_i 的得分，如式(4)所示。

$$H(x,x_i^*,w_i)=\phi(S(x))_i\cdot\Delta P_i^* \tag{4}$$

其中，$\phi(z)_i$ 为 softmax 函数，如式(5)所示。

$$\phi(z)_i=\frac{e^{z_i}}{\sum_{k=1}^{K}e^{z_k}} \tag{5}$$

PWWS 基于 $H(x,x_i^*,w_i)$ 对所有单词降序排列，并选择候选词，贪心遍历整个过程直到使模型标签发生变化，本质上是基于统计的方法。实验结果在 IMDB 等数据集上比 WS 等方法使模型的准确度进一步降低。

Jin 等人[36]提出一种称为 TEXTFOOLER 的黑盒攻击方法，由于 WS 和 PWWS 提出的词显著性方法与 BERT 关注一些单词的统计线索[44]相呼应，作者提出了新的单词替换选择机制。用 I_{w_i} 衡量单词 $w_i\subset x$ 对分类结果 $F(x)=Y$ 的影响并排序。对 I_{w_i} 的计算如式(6)所示。

$$I_{w_i}=\begin{cases}F_Y(x)-F_Y(x^{|w_i}),\\ \quad \text{if } F(x)=F(x^{|w_i})=Y\\ (F_Y(x)-F_Y(x^{|w_i}))+(F_{\bar{Y}}(x^{|w_i})-F_{\bar{Y}}(x)),\\ \quad \text{if } F(x)=Y,\ F(x^{|w_i})=\bar{Y} \text{ and } Y\neq\bar{Y}\end{cases} \tag{6}$$

在单词重要性排序做替换时，先使用余弦相似度选择 N 个近邻作为候选词，再对候选词做词性检查，确保替换后语法基本一致，最后用 USE 方法[45]做语义相似度检查。在筛选后的单词中选择使目标标签发生变化或者使目标标签置信度最低的词做替换，然后重复下一个词。该模型将几乎所有目标模型在所有任务中的精度降低到 10%以下，而只有不到 20%的原始单词受到干扰。文中攻击的对象是文本分类和文本蕴涵任务。

Alzantot 等人[37]首次将群体优化算法引入文本对抗样本攻击中，提出了一种基于遗传算法 GA (genetic algorithm)的单词级黑盒攻击算法，该算法提出了称为 Perturb 的方法作为 baseline，该方法随机选择一个单词，先根据 GloVe 词空间计算该单词的 N 个近邻作为候选词，再利用 Google 语言模型[46]根据上下文筛选最优单词进行排序并保留前 k 个单词，最后从这 k 个词中选择一个使目标标签最大化的词做替换，Perturb 到此结束。随后使用遗传算法做优化，首先调用 Perturb 多次生成初始种群 P^0，初始种群由多个只替换一个单词的新文本构成，如式(7)所示。

$$P^0=(\text{Perturb}(x),\cdots,\text{Perturb}(x)) \tag{7}$$

每一代的个体可以通过查询被攻击模型得到对应目标标签的预测得分，以此预测得分成比例地在第 P 代抽取个体表示为 Sample(P)，如果这些样本中存在使目标标签发生变化的样本，则优化完成，该样本为对抗样本，否则以一定概率成对抽取样本做交叉 Crossover，交叉过程为从两个样本中按顺序随机抽取其中任一文本中的单词，形成新的子代 child。第 $i+1$ 代的第 k 个个体表示如式(8)所示。

$$\text{child}_k^{i+1}=\text{Crossover}(\text{Sample}(P^i),\text{Sample}(P^i)) \tag{8}$$

一轮完成之后再次使用 Perturb 做第二轮优化直到模型标签预测发生变化。实验结果表明，GA 算法在情感分类任务的 IMDB 数据集中，在 25%替换率阈值情况下攻击成功率达到了 97%，在文本蕴含任务中达到了 70%。远高于其提出的 baseline。表 7[37]展示了 GA 算法在情感分类任务生成的对抗样本。

表 7　为情感分类任务生成的对抗样本

Original Text Prediction = **Positive.**
absolutely fantastic whatever I say wouldn't do this underrated movie the justice it deserves watch it now fantastic.
Adversarial Text Prediction = **Negative.**
absolutely fantastic whatever I say wouldn't do this **underestimated** movie the justice it **deserve** watch it now fantastic.

随后 Wang 等人[38]提出了一种改进的遗传算法 IGA(improved genetic algorithm)，与 GA 相比，不同的地方如下：在初始化时，GA 使用 Perturb 随机初始化 S 个样本，而 IGA 采用同义词替换文本中的每一个单词形成文本长度大小的初始种群，使得种群数量更加丰富；在变异时，GA 排除了替换过的单词，而 IGA 可以替换之前替换过的单词，这样可以避免局部最小化，替换后的样本种类更多；在交叉时，GA 是在两个亲代中按顺序随机选择一个单词形成子代，而 IGA 则使用随机剪切，不限于单词，可以剪切文字段。实验结果表明，IGA 攻击后的样本在 Word-CNN、LSTM 和 Bi-LSTM 模型上的测试结果都优于 GA。

谷歌于 2019 年提出了一种黑盒方法来生成包括释义对(paraphrase pairs)的英文数据集 PAWS[39] (paraphrase adversaries from word scrambling)，PAWS 由问答网站 Quora 和维基百科

的句子构成。释义对即含义相同的一对句子，如“Flights from New York to Florida.”和“Flights to Florida from New York.”非释义对为含义相反的一对句子，如“Flights from New York to Florida.”和“Flights from Florida to New York.”可以看出此处的释义对类似于对抗样本，即一对含义相同而又互相有微小差别的句子。目前比较先进的模型，如BERT，仅在现有的自然语言数据集下训练，并不能很好地识别出非释义对，同时会对释义对判断错误，因为现有的数据集缺少诸如此类的释义对和非释义对。在生成释义对时，原始文本首先输入到基于单词扰动(word scrambling)的语言模型，该模型将生成具有单词级交换的扰动文本，但是还无法保证扰动文本与原文本是否为释义对，有时会出现交换单词后语句与原句意义截然相反的情况。为了保持释义和非释义的平衡，算法采用了反向翻译(back translation)加人工判断的形式进一步调整生成的语句。实验结果表明，现有数据集上的模型在PAWS上准确度低于40%，但是，包括了PAWS数据的模型可将其准确度提升到85%，同时能够保持现有任务的性能。

2020年，Zang等人[40]提出了一种新的黑盒攻击方法，该方法结合了基于义原的词替换方法和基于粒子群优化PSO(particle swarm optimization)的搜索方法。基于义原的方法可以保留更多潜在的高质量候选替换词。粒子群算法利用一群相互作用的个体在特定空间中迭代搜索最优解。种群被称为群体，个体被称为粒子，每个粒子在搜索空间中都有一个位置，并以相应的速度运动。算法在初始化种群时采用了遗传算法的方法，后面算法与原始PSO算法相同。该方法基于文本的离散型，不直接更新空间，而是采用是否移动的概率思想向全局最优优化。在SPO的思想上让文本在独立的空间移动。实验表明PSO相比于baseline具有更高的攻击成功率和更高质量的对抗样本。

综上所述，黑盒攻击方法有很多相似之处。由于黑盒不知道模型的内部结构和具体参数，学者们试图通过不同的方法在单词空间中寻找最优替换。有的先替换单词，然后再计算替换后的预测分数，如对遗传算法进行多次迭代寻找最优替换。同时，攻击者也试图找出替换词的重要性分数，然后进行替换，以最小化替换词的比例，如DeepWordBug和TEXTFOOLER。

3 文本对抗样本检测与防御方法

为神经网络生成对抗样本的一个基本目的是利用这些对抗样本增强模型的鲁棒性。事实上，防御比攻击更困难，这方面的工作相较于攻击做得较少。造成这种情况有两个原因：一个是对于复杂的优化问题，如对抗样本生成方法不存在一个好的理论模型；另一个原因是，大量的输入可能产生目标输出的可能性很多，具有很高的不可预知性。因此，建立一个真正的通用性防御方法是困难的。目前文本的对抗样本检测与防御方法还比较少，主要是针对特定模型所提出来的方法。本文将现有的防御方法主要分为以下几类：①改进训练；②添加附加组件。第一种方法不影响模型，第二种方法类似于数据预处理，先对文本处理一遍，再将其输入模型。

3.1 改进训练

3.1.1 对抗训练

对抗训练[6,47]是最早提出来的针对对抗样本的防御方法，其基本思想是使用各种对抗样本生成方法先制作一组大型的对抗样本，并将它们添加到训练数据集中。训练数据集则包括对抗样本和对应的原始样本的混合，通过这种方法可以在一定程度上检测出对抗样本。利用对抗训练可以对深度神经网络进行正则化，减少过拟合，从而提高神经网络的鲁棒性。对抗训练有一些有效的抵御攻击的方法[26,30,48-49]，但对抗训练方法并不总是有效的。2017年，Jia等人[24]通过对抗训练来增强阅读理解模型的鲁棒性，实验结果表明，对抗训练在防御使用相同的生成对抗样本方式的攻击是有效和健壮的。Pruthi等人[50]用对抗训练评估了用于情感分类的BERT模型[51]，当使用字符级替换后的对抗样本攻击BERT时，其准确率由90.3%降低到了64.1%，而使用对抗训练后准确率只恢复到69.2%。

在Alzantot等人的工作[37]中也说明了这一点，其防御方法将生成的1 000个对抗样本加到现有的训练集中，并使用更新的数据集从头开始对模型进行对抗训练。尽管该模型对训练集中包含的对抗样本分类的准确性达到了近100%，但是在使用测试集的实验中没有提供额外的鲁棒性，主要是由于生成对抗样本的方式不同，对抗训练在特定的攻击方式中表现较好，如前面的攻击是通过插入、替换、删除的方式。而Alzantot等人的工作[37]是基于遗传

算法的同义词替换。这也是对抗训练的局限性，它只对相应的攻击有效，对新的攻击方式产生的样本没有效果。

由于生成对抗样本的种类很多，仅靠人工加入的对抗样本做对抗训练来防御各种类型的对抗攻击是不够的。2020 年，Liu 等人[52]提出了一种新的对抗训练方法，该方法的核心思想是迭代训练，自动生成原本没有的对抗样本并重新训练模型，整个过程无须人工干预。具体防御步骤是：对于每次迭代，首先通过扰动词向量训练过程得到权重矩阵 $\boldsymbol{w}$ 和训练损失 trainloss；然后使用贪心搜索的方法从 $\boldsymbol{w}$ 中抽取扰动文本并构造对抗样本；最后在数据集中加入新的对抗样本重新训练模型。重复整个过程直到 trainloss 和迭代次数小于阈值。图 1 展示了算法的流程图，实验结果表明，该方法可以生成更加丰富的对抗样本，从而有效补充了人为添加样本做对抗训练的不足，大大提高了阅读理解模型的鲁棒性。进一步的实验再加上基于规则的方法可以进一步提高模型的鲁棒性，并且超过 SOTA 性能。但是这种方法生成的对抗样本都是不符合语法和没有意义的序列，虽然成功使模型判断错误，但生成对抗样本的方法还有待改进。

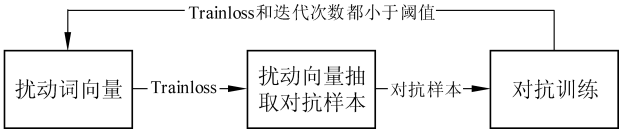

图 1　防御过程流程图

同年 Liu 等人[53]针对字符级攻击提出了联合字符词向量 charCNN 和对抗稳定训练 AST (adversarial stability training)的防御方法。对于字符级攻击，难点在于字符级攻击可能产生词典外的词 OOV (out-of-vocabulary)，即该词不存在，OOV 会带来信息的损失，如何处理好替换词和原词的分布差异是一个难点。同时字符级攻击导致了训练集和对抗样本之间的分布差异，这种差异违反了机器学习中的独立同分布假设，从而导致了模型在对抗样本上性能不佳。字符级词向量不仅可以像单词词向量一样提取语法和语言信息，同时可以提取单词前缀和后缀，尽可能多地保留单词信息。在对抗训练方面采用了字符交换、替换、删除和插入等方式来产生扰动样本。该防御框架有三个损失函数，原文本输入分类器得到的损失 $L_{\text{true}}(x,Y)$，扰动文本输入分类器得到的损失函数 $L_{\text{noisy}}(x',Y)$，还有原文本和扰动文本输入相似度评价模型得到的损失函数 $L_{\text{diff}}(x,x')$，最后得到对抗训练的目标函数，如式(9)所示。

$$\begin{aligned}L(\theta_{\text{cle}},\theta_{\text{cla}})=&\sum_{x\in X}L_{\text{true}}(x,Y,\theta_{\text{clm}},\theta_{\text{cla}})\\&+\alpha\sum_{x'\in X'}L_{\text{noisy}}(x',Y,\theta_{\text{clm}},\theta_{\text{cla}})\\&+\beta\sum_{x'\in X'}L_{\text{diff}}(x,x',\theta_{\text{clm}})\end{aligned}\tag{9}$$

其中，θ_{clm}是字符词向量参数，θ_{cla}是分类器参数，α 和 β 是超参数。在实验结果上，这种联合方法 CharCNN-BLSTM＋AST 在原始文本和对抗样本上分别取得 92.24%和 89.45%，准确率优于 baseline 的 Word-BLSTM 方法的 92.22%和 82.50%。

3.1.2　重新编码

Wang[38]等人提出一种称为同义词编码(SEM)的防御方法，利用句子中的单词是离散标记的特性，可以很容易地找到输入文本的几乎所有邻居。基于此，提出了一种名为 SEM (synonyms encoding method)的新方法来定位输入单词的相邻词。可以假设输入句子 x 的邻居是它的同义句。为了找到同义句，可以用同义词代替句子中的词。要构造这样的映射，需要做的就是对同义词进行整合，并为每个整合后的同义词分配一个唯一的标记。SEM 防御过程是先生成同义词编码字典的词向量矩阵，并将其保存下来；然后使用同义词编码的单词词典训练 CNN、LSTM 和 Bi-LSTM 模型；最后通过攻击方法攻击重新训练的模型。实验结果表明，在受到攻击时，使用 SEM 后的模型正确率高于正常训练和对抗训练的模型。

2020 年，Wang 等人[54]提出随机替换编码 RSE (random substitution encoding)的防御方法，主要攻击过程如下：对于输入的文本 x，在给定的最大替换率和最小替换率之间随机选择一个替换率 s_r；然后产生文本的候选单词集 C，再为 C 中每个单词挑选同义词得到扰动后的文本 x'，用 x'代替 x，最后用这些新的文本训练模型，在测试阶段可以用生成的 x'输入训练好的模型测试 RSE 替换方法对模型的影响。实验结果表明，在 CNN、LSTM、Bi-LSTM 三种模型中，IMDB、AGs News 和 Yahoo! Answers 三种数据集上，RSE 的防御效果整体优于 SEM。

同年，Erik 等人[55]提出了一种称为 RobEn (robust encodings)的防御方法，RobEn 针对字符级攻击，其核心是一种编码功能，将句子映射到较小

的离散编码空间，然后将其用于预测。RobEn的思想类似于SEM，主要区别是它在同义词的选取上是字符级修改后的单词，同时考虑了OOV词的影响，而SEM为单词级别上的相似词。Erik等人指出了重新编码方法防御对抗样本攻击需要满足的两个方面：一是稳定性，为了保证不同扰动的一致预测，编码需要将句子的所有扰动映射到一组编码上；二是保真性，需要保证使用重新编码训练的模型在原始未受扰动的输入上仍然效果良好，即重新训练的模型准确率不应低于原始训练的模型。而SEM在保真性上表现不佳，RobEn的编码函数如式(10)所示。

$$\alpha(x)=[\pi(x_1),\pi(x_2),\cdots,\pi(x_L)] \quad (10)$$

将每个单词x_i映射为$\pi(x_i)$，将许多单词及其错别字映射到相同的编码token。实验结果表明，RobEn在恢复准确率上比单词拼写检查方法高36%。

3.2 添加附加组件

对抗攻击的一种防御策略是检测输入数据是否被修改。对抗样本和非对抗样本之间由于修改过肯定存在一些不同的特征。根据这一点，人们进行了一系列的工作来检测对抗样本，并在图像上[56-57]表现得相对较好。在文本中，某些方法的修改策略可能会在生成的对抗样本中产生拼写错误，这是一个可以利用的特性。Pruthi等人[50]通过检查拼写错误的单词来检测对抗样本，该方法在分类器的前面放置一个单词识别模型。单词识别模型建立在RNN半字符体系结构的基础上，引入了一些新的策略来处理稀有的单词。经过训练，该模型可以识别由随机添加、删除、交换和键盘错误而改变的单词，相对于原始半字符模型，该方法减少了32%模型对对抗样本的错误分类。Li等人[48]利用上下文感知拼写检查服务来做类似的工作。实验结果表明，该方法对字符级的修改是有效的，但不适用于单词级的修改，因为单词级的替换通常是真实存在的单词，没有拼写上的错误。

Zhou等人[58]采用对模型添加部件的方式来防御对抗攻击，主要用来防御单词级分类器。该方法有三个组件：首先是扰动检测器，先检测出一些候选的可能被扰动单词；然后是词估计器，利用BERT等方法抽取给定上下文特征，找出最优的可能原始词；最后是单词水平的还原。之后，将全部恢复的文本输入给模型进行预测。如能恢复为原始标签，则防御成功。这种方法称为DISP(discriminate perturbations)。实验结果表明，DISP在单词级攻击的防御效果优于对抗训练[47]和拼写检查[50]，同时将DISP与拼写检查结合能够进一步提高模型鲁棒性。

4 中文对抗样本攻击与防御

与英文对抗样本不同，大多数需要解决的中文对抗性文本是由现实世界中的恶意网民生成的，由于不同网民采用了不同修改策略，因此其多样性更加丰富，如生成垃圾邮件逃避检测[59]、在线论坛发布广告等恶意内容。由于中文比较好的词向量预训练模型还较少，这使得现有的工作在同义词替换上还存在一些不足。另外，中文中的字符空间非常大，每个字符都可能受到各种攻击策略的干扰，这使得干扰更加稀疏和无规律可循。目前针对中文的对抗样本攻击和防御的研究较少，本文各选取了一种效果较好的攻击和防御方法来说明对抗样本在中文中的应用。

4.1 针对中文的对抗样本攻击

王文琦等人[60]提出了一种称为WordHanding的黑盒攻击方法，这种方法设计了新的中文词重要性计算方法，使用同音词替换做对抗攻击。与WS、PWWS等方法类似，WordHanding定义了一些新的衡量方法来计算词重要性，最后得到词重要性的排序，再做词替换。实验数据集采用京东购物评论和携程酒店评论，结果表明WordHanding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%。同音词替换虽然可以很好地表达原始文本的意思，因为读上去的发音跟原始文本一样，但是由于只是读音相同，在视觉上能够明显区分出这种对抗样本，文本可读性较差。表8[60]展示了WordHanding生成的对抗样本。可以看到，对抗样本替换的词都是感情色彩比较明显的词，对情感分类影响较大。

表8 原始文本和WordHanding生成的对抗样本

原始文本：屏幕较差，拍照也很粗糙.
对抗样本：屏幕交叉，拍照也很**出操**.
原始文本：服务态度不好，换个房间都不给换，弄个最差的给住.
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.

4.2　中文对抗样本攻击的防御

2014 年，有研究者提出了针对中文的拼写检查和纠错的方法，包括基于规则[61]的和基于语言模型[62]的方法。但是相比于英文，中文拼写检查更困难，因为中文文本词语之间没有分割，一句话是连在一起的整体，而拼写检测一般只能在词级别上来确定，结果是中文的拼写检测对模型性能的恢复能力有限。

Li 等人[63]提出了一种专门为基于中文对抗样本的防御框架 TEXTSHIELD，这是一种基于神经机器翻译（NMT）的新型防御方法。首先在一个大型对抗性平行语料库上训练一个 NMT 模型，基于 Encoder-Decoder 框架设计对抗 NMT 模型，用于恢复扰动的文本；然后将其置于基于多模态嵌入的文本分类模型前，将对抗性扰动恢复到良性扰动。在训练阶段，首先通过对抗性攻击生成大量的句子对，构造一个大型对抗性平行语料库；接下来将设计好的 NMT 模型在对抗样本上进行训练。一旦训练完成，NMT 模型就会被用作对抗文本校正器，通过翻译来重建原始文本。由于攻击者在真实场景中采用的扰动策略主要集中在基于符号和基于语音的扰动上，该文专门提出了三种跨不同模式的词向量方法来处理相应的变异类型：即语义词向量：应用 skip-gram 模型[64]来学习连续的语义词向量。字形词向量：在中国的书写体系中，有一大批文字在视觉上很相似，但却有着完全不同的含义，例如，用“堵”取代“赌博”的“赌”，用字形词向量方案来提取每个字符的基于字形的特征，以捕捉受干扰的单词与其良性对应词之间的相似性。语音词向量：提取汉字中的同音字，提高文本分类模型及其鲁棒性，例如扰动如“涩情”或“se 清”，和毒性词“色情”有相同的发音。实验结果表明，TEXTSHIELD 对恶意用户生成的对抗文本具有较高的准确性，如对“色情”检测准确率为 0.944，而与良性输入相比，对模型性能的影响很小，模型准确性降低了不到 2%。

5　面临的挑战和前景展望

5.1　单词相似性与预训练模型使用问题

对抗样本的研究由图像发展而来，由于图像具有连续性，使得图像的对抗样本有着很好的人类不可感知性。但是，在文本领域，一点很小改动都能明显被人类察觉到，虽然现在主流的方法是同义词替换，但这种同义词是词空间中由词向量计算而来的同义词，且词向量中距离较近的同义词不一定是现实生活中的同义词，在有些对抗样本攻击中生成的同义词在现实中是不相关词，在不同的数据集中，一个词的同义词可能因为语境等原因有差别。同时，单词的替换很可能改变句子的语义或语法，虽然最新的攻击方法在尽量避免这种情况，但仍然有很大的提升空间。因此，使用能准确反映训练数据集单词相似性的词向量将会大大改善生成的对抗样本质量。如在中文同义词替换部分使用 ERNIE[65]（enhanced representation through knowledge integration）等其他表现较好中文预训练语言模型，ERNIE 的训练语料库包括了维基百科、百度贴吧、百度新闻等常见的中文表现形式。其中百度贴吧由于其 AI 文本审核技术的存在，导致百度贴吧本身就存在很多规避审核的对抗样本，ERNIE 无论在攻击领域还是防御领域都可以用在未来的工作中。在英文方面，能够更好地反映上下文关系的预训练语言模型 ELECTRA[66]（efficiently learning an encoder that classifies token replacements accurately）在训练过程中由流行的 MLM（masked language modeling）方法改进为 RTD（replaced token detection），其将生成式的 MLM 任务改成了判别式的 RTD 任务，其替换词检测的思想跟对抗样本检测和防御非常契合，在未来的工作中使用类似的预训练语言模型或许有助于提升防御效果。

5.2　防御方法的通用性不强

到目前为止，现有的防御方法都是针对特定攻击方法、特定数据集，比如拼写检查在字符级攻击的防御效果较好，但不适用于单词级攻击。特定的防御方法在特定的攻击方法上表现很好，但是换了另一种攻击方法就达不到预期的效果。比如通过分析数据发现 RSE 在 PWWS 攻击方法上防御效果很好，但在 GA 攻击下并不能很好地感知出其所替换的单词。原因是在做防御时，可供选择的攻击方法有限，一般只选择几种方法做实验，算法可以在这几种攻击方法上取得较好的防御效果，但遇到未使用过的攻击方法时防御效果就减弱了。由于攻击方法层出不穷，通用性问题目前还没有较好的解决方法，希望未来有更好的解决方法。

5.3　评估方法

对抗样本攻击效果是一个多因素综合考虑的结

果，现在有一些对抗样本攻击方法以模型准确性来评价其方法的性能，即准确性越低，攻击效果越好。也有一些以攻击成功率评价攻击效果，即成功使模型出错的样本数量与总的样本的比例。但是模型准确性或者攻击成功率只反映了一个方面，另外的比如语义相似性、单词替换率和不可感知性等因素都会影响最后的攻击效果。一个攻击方法的语义相似越高，单词替换率越低，不可感知性越强，其效果越优，但严格限制这几个因素之后攻击成功率就不会太高。对于语义相似性，评估的方法有欧氏距离、余弦距离和编辑距离等方法，这几种方法都有人使用过，但是没有一个标准方法来衡量语义相似性。其次，针对不可感知性，目前通用的方法是人工评估，即选择若干志愿者人工评估生成的对抗样本是否是修改过的对抗样本。严格来说，这并不严谨，但目前还没有更好的评估方法，因此需要更标准化的方法来评估对抗样本的不可感知性和语义相似性。

6 总结

本文从攻击和防御两方面在对抗样本这一新兴的研究方向进行综述，介绍了目前常见的攻击和防御方法，在攻击部分按照黑盒、白盒、字符级和单词级详细介绍了具体的攻击方法，并对所有介绍的攻击方法做了对比。本文还讨论了防御方法，并探讨了现有防御方法的不足。同时介绍了中文对抗样本攻击与防御方法的进展，虽然现有的针对中文的方法较少且不成熟，但在对抗样本这一邻域的研究越来越多，以后针对中文的方法也会更加完善。虽然对抗样本是深度学习的一个重大威胁，但相信随着对于对抗攻击方法的深入研究，未来一定能够实现抵御对抗攻击更具鲁棒性的深度学习模型。

参考文献

[1] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks [J]. Communications of the ACM, 2017, 60(6): 84-90.

[2] O'Mahony N, Campbell S, Carvalho A, et al. Deep learning vs traditional computer vision[C]//Proceedings of the Science and Information Conference. Springer, Cham, 2019: 128-144.

[3] Wang F, Jiang M, Qian C, et al. Residual attention network for image classification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 3156-3164.

[4] Belinkov Y, Glass J. Analysis methods in neural language processing: A survey[J]. Transactions of the Association for Computational Linguistics, 2019, 7: 49-72.

[5] Otter D W, Medina J R, Kalita J K. A survey of the usages of deep learning for natural language processing [J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 32(2): 604-624.

[6] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks[C]//Proceedings of the International Conference on Learning Representations. 2014.

[7] Guo X, Zhu E, Yin J. A fast and accurate method for detecting fingerprint reference point[J]. Neural Computing and Applications, 2018, 29(1): 21-31.

[8] Chen D, Bolton J, Manning C D. A thorough examination of the CNN/Dailymail reading comprehension task [C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 2016: 2358-2367.

[9] Su D, Zhang H, Chen H, et al. Isrobustness the cost of accuracy? --a comprehensive study on the robustness of 18 deep image classification models[C]//Proceedings of the European Conference on Computer Vision. 2018: 631-648.

[10] Wang G, Li C, Wang W, et al. Joint embedding of words and labels for text classification[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018: 2321-2331.

[11] Xue W, Li T. Aspect based sentiment analysis with gated convolutional networks[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018: 2514-2523.

[12] Wang Z, Hamza W, Florian R. Bilateral multi-perspective matching for natural language sentences [C]//Proceedings of the 26th International Joint Conference on Artificial Intelligence, 2017: 4144-4150.

[13] Bastings J, Titov I, Aziz W, et al. Graphconvolutional encoders for syntax-aware neural machine translation[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2017: 1957-1967.

[14] Henderson P, Sinha K, Angelard-Gontier N, et al. Ethical challenges in data-driven dialogue systems [C]//Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, 2018: 123-129.

[15] Liang B, Li H, Su M, et al. Deep text classification can be fooled[C]//Proceedings of the 27th International Joint Conference on Artificial Intelligence,

2018：4208-4215.

[16] 王春柳，杨永辉，邓霏，等.文本相似度计算方法研究综述[J].情报科学，2019，37(03)：158-168.

[17] Kusner M，Sun Y，Kolkin N，et al. From word embeddings to document distances[C]//Proceedings of the International Conference on Machine Learning，2015：957-966.

[18] Rubner Y，Tomasi C，Guibas L J. A metric for distributions with applications to image databases[C]//Proceedings of the 6th International Conference on Computer Vision，1998：59-66.

[19] Maas A，Daly R E，Pham P T，et al. Learning word vectors for sentiment analysis[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics，2011：142-150.

[20] Zhang X，Zhao J，LeCun Y. Character-level convolutional networks for text classification[J]. Advances in Neural Information Processing Systems，2015，28：649-657.

[21] Bowman S，Angeli G，Potts C，et al. A large annotated corpus for learning natural language inference[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing，2015：632-642.

[22] Zhang W E，Sheng Q Z，Alhazmi A，et al. Adversarial attacks on deep-learning models in natural language processing：A survey[J]. ACM Transactions on Intelligent Systems and Technology，2020，11(3)：1-41.

[23] Alshemali B，Kalita J. Improving the reliability of deep neural networks in NLP：A review[J]. Knowledge-Based Systems，2020，191：105210.

[24] Jia R，Liang P. Adversarial examples for evaluating reading comprehension systems[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing，2017：2021-2031.

[25] 孟东宇. 黑盒威胁模型下深度学习对抗样本的生成[J]. 电子设计工程，2018(24)：35.

[26] 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，2018：31-36.

[27] Ebrahimi J，Lowd D，Dou D. On adversarial examples for character-level neural machine translation[C]//Proceedings of the 27th International Conference on Computational Linguistics，2018：653-663.

[28] Tsai Y T，Yang M C，Chen H Y. Adversarial attack on sentiment classification[C]//Proceedings of the 2019 ACL Workshop BlackboxNLP：Analyzing and Interpreting Neural Networks for NLP，2019：233-240.

[29] 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，2019，7：387-401.

[30] Gao J，Lanchantin J，Soffa M L，et al. Black-box generation of adversarial text sequences to evade deep learning classifiers[C]//Proceedings of the IEEE Security and Privacy Workshops (SPW)，2018：50-56.

[31] Heigold G，Varanasi S，Neumann G，et al. How robust are character-based word embeddings in tagging and MT againstword scrambling or random noise?[C]//Proceedings of the 13th Conference of the Association for Machine Translation in the Americas. 2018：68-80.

[32] Eger S，Şahin G G，Rücklé A，et al. Text processing like humans do：visually attacking and shielding NLP systems[C]//Proceedings of NAACL-HLT，2019：1634-1647.

[33] Samanta S，Mehta S. Towards crafting text adversarial samples[J]. arXiv preprint arXiv：1707.02812，2017.

[34] Gan W C，Ng H T. Improving the robustness of question answering systems to question paraphrasing[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics，2019：6065-6075.

[35] 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，2019：1085-1097.

[36] Jin D，Jin Z，Zhou J T，et al. IsBert really robust? a strong baseline for natural language attack on text classification and entailment[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2020，34(05)：8018-8025.

[37] Alzantot M，Sharma Y，Elgohary A，et al. Generating natural language adversarial examples[C]//Proceedings of the 2018 Conference on Empirical Methods in Natura，Language Processing. 2018：2890-2896.

[38] Wang X，Jin H，He K. Natural language adversarial attacks and defenses in word level[J]. arXiv preprint arXiv：1909.06723，2019.

[39] 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，2019：1298-1308.

[40] Zang Y，Qi F，Yang C，et al. Word-level textual ad-

versarial attacking as combinatorial optimization [C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020: 6066-6080.

[41] Kim Y, Jernite Y, Sontag D, et al. Character-Aware neural language models[C]//Proceedings of the 30th AAAI Conference on Artificial Intelligence, 2016: 2741-2749.

[42] Belinkov Y, Bisk Y. Synthetic and natural noise both break neural machine translation[J]. arXiv preprint arXiv: 1711.02173, 2017.

[43] Behjati M, Moosavi-Dezfooli S M, Baghshah M S, et al. Universal adversarial attacks on text classifiers [C]//Proceedings of 2019 IEEE International Conference on Acoustics, Speech and Signal Processing. 2019: 7345-7349.

[44] Niven T, Kao H Y. Probing neural network comprehension of natural language arguments[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 4658-4664.

[45] Cer D, Yang Y, Kong S, et al. Universal sentence encoder[J]. arXiv preprint arXiv: 1803.11175, 2018.

[46] Chelba C, Mikolov T, Schuster M, et al. One billion word benchmark for measuring progress in statistical language modeling[J]. arXiv preprint arXiv: 1312.3005, 2013.

[47] Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples[J]. arXiv preprint arXiv: 1412.6572, 2014.

[48] Li J, Ji S, Du T, et al. Textbugger: Generating adversarial text against real-world applications[J]. arXiv preprint arXiv: 1812.05271, 2018.

[49] Zhao S, Cai Z, Chen H, et al. Adversarial training based lattice LSTM for Chinese clinical named entity recognition[J]. Journal of Biomedical Informatics, 2019, 99: 103290.

[50] Pruthi D, Dhingra B, Lipton Z C. Combating adversarial misspellings with robust word recognition [C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 5582-5591.

[51] Devlin J, Chang M W, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics. 2019: 4171-4186.

[52] Liu K, Liu X, Yang A, et al. A robust adversarial training approach to machine reading comprehension [C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(05): 8392-8400.

[53] Liu H, Zhang Y, Wang Y, et al. Joint character-level word embedding and adversarial stability training to defend adversarial text[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(05): 8384-8391.

[54] Wang Z, Wang H. Defense of word-level adversarial attacks via random substitution encoding[C]//Proceedings of the International Conference on Knowledge Science, Engineering and Management. Springer, Cham, 2020: 312-324.

[55] Jones E, Jia R, Raghunathan A, et al. Robust encodings: A framework for combating adversarial typos [C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 2752-2765.

[56] Xu W, Evans D, Qi Y. Feature squeezing: Detecting adversarial examples in deep neural networks[J]. arXiv preprint arXiv: 1704.01155, 2017.

[57] Roth K, Kilcher Y, Hofmann T. The odds are odd: a statistical test for detecting adversarial examples [C]//Proceedings of the International Conference on Machine Learning. 2019: 5498-5507.

[58] Zhou Y, Jiang J Y, Chang K W, et al. Learning to-discriminate perturbations for blocking adversarial attacks in text classification[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. 2019: 4906-4915.

[59] Jiang Z, Gao Z, He G, et al. Detect camouflaged spam content via Stone Skipping: graph and text joint embedding for Chinese character variation representation[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, 2019: 6188-6197.

[60] 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.软件学报, 2019, 30(8): 2415-2427.

[61] Yeh J F, Lu Y Y, Lee C H, et al. Chinese word spelling correction based on rule induction[C]//Proceedings of the 3rd CIPS-SIGHAN Joint Conference on Chinese Language Processing, 2014: 139-145.

[62] Yu J, Li Z. Chinese spelling error detection and correction based on language model, pronunciation, and shape[C]//Proceedings of the 3rd CIPS-SIGHAN Joint Conference on Chinese Language Processing, 2014: 220-223.

[63] Li J, Du T, Ji S, et al. TextShield: robust text classification based on multimodal embedding and neural machine translation [C]//Proceedings of the29th USENIX Security Symposium, 2020: 1381-1398.

[64] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[C]//Proceedings of the Advances in Neural Information Processing Systems, 2013: 3111-3119.

[65] Sun Y, Wang S, Li Y, et al. Ernie: Enhanced representation through knowledge integration[J]. arXiv preprint arXiv: 1904.09223, 2019.

[66] Clark K, Luong M T, Le Q V, et al. ELECTRA: Pre-training text encoders as discriminators rather than generators[C]//Proceedings of the International Conference on Learning Representations, 2019.

杜小虎(1994—)，硕士研究生，主要研究领域为自然语言处理、人工智能安全。
E-mail: duxiaohu18@nudt.edu.cn

吴宏明(1983—)，硕士，工程师，主要研究领域为人工智能、军事大数据。
E-mail: 289326324@qq.com

易子博(1991—)，博士研究生，主要研究领域为人工智能、对抗样本。
E-mail: yizibo14@nudt.edu.cn

庆祝中国中文信息学会成立 40 周年系列活动—第十六届中国中文信息学会暑期学校暨《前沿技术讲习班》(CIPS ATT)在京召开

2021 年 7 月 22 日-25 日，庆祝中国中文信息学会成立 40 周年系列活动—第十六届中国中文信息学会暑期学校暨《前沿技术讲习班》第二十三期和第二十四期在京举行。本届讲习班主题为：预训练语言模型的基础理论与方法及其典型应用。讲习班吸引了来自全国各高校及科研院所的专家、学者、学生、产业界研发人员等近 400 人参加。

哈尔滨工业大学车万翔教授、清华大学刘知远副教授、中国科学院自动化研究所张家俊研究员担任本届讲习班的学术主席。刘知远副教授和车万翔教授分别致开幕辞，先后介绍了讲习班的课程内容和特邀讲者，并欢迎学员们来京参会！

讲习班邀请了哈尔滨工业大学车万翔教授、科大讯飞崔一鸣研究员、微软亚洲研究院董力研究员、百度公司孙宇研究员、清华大学刘知远副教授、循环智能杨植麟博士、复旦大学邱锡鹏教授、上海交通大学赵海教授、字节跳动王明轩研究员、清华大学黄民烈副教授、中国人民大学赵鑫副教授和陈旭助理教授、中科院计算所郭嘉丰教授和范意兴助理研究员、清华大学兰艳艳教授、中科院自动化所刘康研究员和中科院软件所韩先培研究员从不同的方向作了系统深入的讲解，并对预训练模型及其各领域的应用提出了未来的研究目标。