面向文本分类的类别区分式通用对抗攻击方法

郝志荣，陈龙，黄嘉成

引用本文

郝志荣，陈龙，黄嘉成. 面向文本分类的类别区分式通用对抗攻击方法[J]. 计算机科学, 2022, 49(8): 323-329.

HAO Zhi-rong, CHEN Long, HUANG Jia-cheng. Class Discriminative Universal Adversarial Attack for Text Classification[J]. Computer Science, 2022, 49(8): 323-329.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220200077

面向文本分类的类别区分式通用对抗攻击方法

郝志荣[1] 陈 龙[1,2] 黄嘉成[1]
1 重庆邮电大学计算机科学与技术学院 重庆 400065
2 重庆邮电大学网络空间安全与信息法学院 重庆 400065
(s190201031@stu.cqupt.edu.cn)

摘 要 通用对抗攻击只需向任意输入添加一个固定的扰动序列,就可以成功混淆文本分类器,但是其会不加区分地攻击所有类别的文本样本,容易引起防御系统的注意。为了实现攻击的隐蔽性,文中提出了一种简单高效的类别区分式通用对抗攻击方法,突出对目标类别的文本样本有攻击效果,并尽量对非目标类别不产生影响。在白盒攻击的场景下,利用扰动序列在每个批次上的平均梯度搜索得到多个候选扰动序列,选择损失最小的扰动序列进行下一轮迭代,直到没有新的扰动序列产生。在4个公开的中英文数据集以及神经网络模型TextCNN和BiLSTM上进行了大量的实验,以评估所提方法的有效性,实验结果表明,该攻击方法可以实现对目标类别和非目标类别的区分式攻击,而且具有一定的迁移性。

关键词:通用对抗攻击;文本分类;类别区分式;深度学习;神经网络
中图法分类号 TP183

Class Discriminative Universal Adversarial Attack for Text Classification

HAO Zhi-rong[1],CHEN Long[1,2] and HUANG Jia-cheng[1]
1 School of Computer Science and Technology,Chongqing University of Posts and Telecommunications,Chongqing 400065,China
2 School of Cyber Security and Information Law,Chongqing University of Posts and Telecommunications,Chongqing 400065,China

Abstract The definition of universal adversarial attack is that the text classifiers can be successfully fooled by a fixed sequence of perturbations appended to any inputs. But textual examples from all classes are indiscriminately attacked by the existing UAA, which is easy to attract the attention of the defense system. For more stealth attack,a simple and efficient class discriminative universal adversarial attack method is proposed,which has an obvious attack effect on textual examples from the targeted classes and limited influence on the non-targeted classes. In the case of white-box attack, multiple candidate perturbation sequences are searched by using the average gradient of the perturbation sequence in each batch. The perturbation sequence with the smallest loss is selected for the next iteration until no new perturbation sequence is generated. Comprehensive experiments are conducted on four public Chinese and English datasets and TextCNN,BiLSTM to evaluate the effectiveness of the proposed method. Experimental results show that the proposed attack method can discriminatively attack the targeted and non-targeted classes,and has certain transferability.

Keywords Universal adversarial attack,Text classification,Class discriminative,Deep learning,Neural Networks

1 引言

近年来,深度神经网络在自然语言处理领域的成效十分显著。文本分类作为一个基础任务,在现实中的应用十分广泛,如情感分析、新闻分类、垃圾邮件分类等。这些基于深度神经网络的文本分类器在遇到对抗攻击[1-3]时,性能会显著下降,引起了人们对应用安全性、有效性的担忧。同时,了解对抗攻击不仅有助于评估模型的能力[4],还可以在一定程度上解释模型的脆弱性[5-6],因此得到了学术界和工业界的广泛关注和深入研究。

文本对抗样本可以通过对原始文本的某些字符、词、短语进行若干次插入、替换、交换和删除等扰动操作而生成。对抗攻击指生成的对抗样本可以愚弄文本分类器,但不影响人们对其原始语义的理解。按照扰动操作是否一样可以将对抗攻击分为两种:一种是样本相关性对抗攻击[7],每个对抗样本的扰动操作会随着原始文本的变化而变化;另一种是通用对抗攻击(Universal Adversarial Attack,UAA)[8-10],所有对抗样本的扰动操作都是添加一个固定的扰动序列。UAA可以

到稿日期:2022-02-15 返修日期:2022-03-24
基金项目:重庆市教委重点合作项目(HZ2021008)
This work was supported by the Key Cooperation Project of Chongqing Municipal Education Commission(HZ2021008).
通信作者:陈龙(chenlong@cqupt.edu.cn)

提前计算扰动序列并直接将其应用于实际攻击，而且生成的对抗样本都包含相同的扰动序列。与样本相关性对抗攻击相比，UAA 虽然花费的攻击成本更低，但更容易被防御系统检测到。

为了生成隐藏性更强的扰动序列，本文提出的 CD-UAA（Class Discriminative UAA，CD-UAA）缩小了 UAA 的攻击范围，即文本分类器在处理目标类别的文本时，分类性能会显著下降，而在处理非目标类别的文本时，性能只受轻微影响。CD-UAA 的概述如图 1 所示，目标模型是 BiLSTM，将数据集 ag_news 的 Business 选为目标类别，将其余类别（World，Sports，Sci/Tech）选为非目标类别，“xss browsers hariri”是生成的扰动序列，将其添加到测试集所有文本样本的前面以构成对抗样本，模型 BiLSTM 对对抗样本表现出了不同的性能变化。为了限制扰动序列对非目标类别的影响，CD-UAA 首先从训练集中随机采样多个批次的文本样本，并初始化一个扰动序列，将其添加到所有文本样本的前面，然后对目标和非目标类别的文本样本使用不同的损失函数计算其损失，最后利用扰动序列在所有批次上的梯度信息中迭代搜索损失最小的扰动序列。本文的主要贡献如下：

（1）提出了一种针对文本分类器的类别区分式通用对抗攻击方法（CD-UAA），并在 2 个模型和 4 个数据集的多个组合上评估了其攻击性能。

（2）CD-UAA 仅使用 32 条训练数据就可以生成攻击效果较好的扰动序列，而且利用多个不同的损失函数配置可以进一步提升攻击效果。

（3）对 CD-UAA 进行了全面的评估，包括在不同情况下与 UAA 的对比、超参数和扰动序列长度对其的影响以及扰动序列在不同模型之间的迁移性。

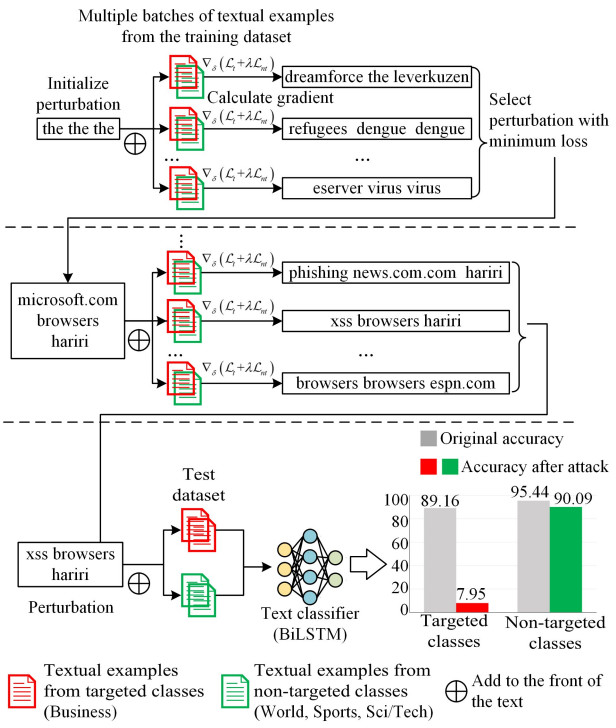


图 1 CD-UAA 对 BiLSTM 性能的影响

Fig. 1 Influence of CD-UAA on the performance of BiLSTM

2 相关工作

Moosavi-Dezfooli 等[11]首先在图像领域提出了 UAA 的概念，随后 Behjati 等[8]利用梯度投影实现了针对文本分类器的 UAA。而 Wallace 等[9]提出了一种基于梯度引导的 UAA，并将其应用于文本分类、阅读理解和文本生成等任务，UAA 被认为是一种独特的模型分析工具。Song 等[10]在 UAA 的基础上利用逆向正则化自动编码器（Adversarially Regularized Autoencoder，ARAE）保证了扰动序列在语义上的流畅性。Heidenreich 等[12]则发现，UAA 可以轻松控制条件文本生成模型（如 GPT-2）生成内容的主题。

由于 UAA 不能让攻击者灵活地控制攻击的类别，因此 Gupta 等[13]利用神经网络权重的线性函数实现了目标类别的通用对抗攻击，但未考虑生成的扰动对非目标类别的影响；Zhang 等[14]通过设计分离的损失函数，在保证攻击效果的同时，尽可能地减小对非目标类别的影响；随后 Benz 等[15]在其基础上重新设计了损失函数，并实现了双目标的 UAA，即生成的扰动可以将特定类别的样本攻击为另一种类别，同时限制对非目标类别的影响。由于文本的离散性，图像领域的对抗攻击方法并不能直接迁移到文本领域。

3 类别区分式通用对抗攻击方法

3.1 问题定义

定义 1 给定一个 m 分类的文本分类模型 $F:\{X\to Y\}$；由词或字符组成的文本样本集合 $x\sim X$，其中 X 服从数据分布 $D(X)$；类别集合 $y\sim Y$。在 CD-UAA 中，选定需要攻击的目标类别集合 S，按照样本类别是否属于 S，将 X 分为目标类别的文本样本集合 $x_t\sim X_t$ 和非目标类别的文本样本集合 $x_{nt}\sim X_{nt}$，找到一个长度为 k 的扰动序列 $\delta=\{\delta_1,\delta_2,\cdots,\delta_k\}$，将其插入到文本样本 x_t 或 x_{nt} 的某一位置，这样可以成功混淆模型 F 对于 X_t 的输出，同时减小对 X_{nt} 的影响，其中 k 不应太长，以免影响对抗样本的原有含义。CD-UAA 的攻击目标可以表示为：

$$\begin{cases}F(\delta,x_t)\neq F(x_t), & \text{for most } x_t\sim X_t\\ F(\delta,x_{nt})=F(x_{nt}), & \text{for most } x_{nt}\sim X_{nt}\end{cases}\tag{1}$$

3.2 攻击方法描述

在文本分类器的白盒状态下，以 Wallace 等[9]提出的 UAA 为基础，针对特定的目标类别，考虑如何生成一个 CD-UAA 的扰动序列。首先确定攻击的目标类别集合 S 和扰动序列的长度 k。初始化扰动序列，其过程大致分为两种：一种是通过重复词汇表中无意义的词，如“的”或“the”；另一种是从词汇表中随机选择一些词。

然后，分别从 X_t 和 X_{nt} 中随机取样多个批次的文本，用于搜索扰动序列，在每个批次中，目标类别与非目标类别的文本数量各占一半。由于文本的离散性，在 HotFlip[16]算法的基础上利用扰动序列的梯度来近似表示其扰动效果。将扰动序列 $\delta=\{\delta_1,\delta_2,\cdots,\delta_k\}$ 表示成 One-hot 向量 $\boldsymbol{e}_\delta$，利用式（2）在模型词汇表的嵌入向量空间中搜索一个扰动序列，使当前批次的损失最小。

$$\arg\min_{e\in\mathcal{V}}[\boldsymbol{e}-\boldsymbol{e}_\delta]^{\mathrm{T}}\nabla\boldsymbol{e}_\delta\mathcal{L} \tag{2}$$

其中，$\mathcal{V}$是模型词汇表的嵌入向量空间，$\boldsymbol{e}$是嵌入向量空间的一个词向量；$\nabla\boldsymbol{e}_\delta\mathcal{L}$是损失函数$\mathcal{L}$关于扰动序列$\delta$在一批文本数据的平均梯度。因此，针对所有批次的训练数据，都可以利用扰动序列δ在当前批次的平均梯度搜索得到一个候选扰动序列。训练数据越多，生成的候选扰动序列也就越多，最后从中筛选出一个扰动序列，以使所有批次的平均损失最小，将其替换为下一轮迭代的扰动序列，直到扰动序列不再更新。

除了搜索扰动序列的过程，损失函数对攻击效果也有十分重要的影响。Zhang 等[14]已经证实，设计两个损失函数$\mathcal{L}_t$和$\mathcal{L}_{nt}$来分别处理目标类别以及非目标类别的样本，对于 CD-UAA 来说是合理且有效的，损失函数$\mathcal{L}$的优化方向与$\mathcal{L}_t$一致，表达式如下：

$$\mathcal{L}=\mathcal{L}_t(x_t)+\lambda\mathcal{L}_{nt}(x_{nt}),\ \text{for}\ x_t\sim X_t, x_{nt}\sim X_{nt} \tag{3}$$

其中，λ是影响$\mathcal{L}_{nt}$权重的超参数，可以让攻击者控制$\mathcal{L}_{nt}$对攻击效果的影响，即λ越大，对X_{nt}的攻击成功率越小。为了方便描述，$L_i(\cdot)$表示文本分类模型第i个类别的 logit 值；$c=\arg\max F(x)$表示原始文本的预测分类；$ReLu(x)=\max(x,0)$；CE表示交叉熵损失函数。损失函数$\mathcal{L}_t$处理目标类别的文本，应该降低预测类别的 logit 值，其有以下 3 种具体表示：

$$\begin{cases}\mathcal{L}_t^{CE}=-CE(F(x_\delta,x_t),F(x_t))\\ \mathcal{L}_t^{LC}=L_c(x_\delta,x_t)\\ \mathcal{L}_t^{BL}=\mathrm{ReLu}(L_c(x_\delta,x_t)-\max\limits_{i\neq c}L_i(x_\delta,x_t))\end{cases} \tag{4}$$

而损失函数$\mathcal{L}_{nt}$处理非目标类别的文本，应该增加预测类别的 logit 值，使其保持最高，其有以下 3 种具体表示：

$$\begin{cases}\mathcal{L}_{nt}^{CE}=CE(F(x_\delta,x_{nt}),F(x_{nt}))\\ \mathcal{L}_{nt}^{LC}=-L_c(x_\delta,x_{nt})\\ \mathcal{L}_{nt}^{BL}=\mathrm{ReLu}(\max\limits_{i\neq c}L_i(x_\delta,x_{nt})-L_c(x_\delta,x_{nt}))\end{cases} \tag{5}$$

$\mathcal{L}_t^{CE}$和$\mathcal{L}_{nt}^{CE}$考虑了所有类别的 logit，可能对另一损失函数的影响较大。$\mathcal{L}_t^{LC}$和$\mathcal{L}_{nt}^{LC}$只考虑了预测类别的 logit，可能对另一损失函数的影响较小，而$\mathcal{L}_t^{BL}$和$\mathcal{L}_{nt}^{BL}$通过限制其他类别的 logit 来进行优化，并通过 ReLu 函数限制优化边界。CD-UAA 的具体步骤如算法 1 所示。

算法 1 CD-UAA

输入：数据分布 X；文本分类模型 F；嵌入向量空间 **V**；最大迭代次数 MaxStep

输出：扰动序列 δ

```
1. X_t, X_nt ⊆ X
2. 初始化扰动序列 δ
3. for i=1 to MaxStep do
4.     Arrays=[δ]
5.     for j, b_t, b_nt ← X_t, X_nt do
6.         avg_δ ← ∇_δ L_(b_t, b_nt) //计算平均梯度
7.         δ_j ← arg min_{e∈V} [e−e_δ]^T avg_δ //候选扰动序列
8.         Arrays.append(δ_j)
9.     end for
10.    δ_min ← arg min_{a∈Arrays} AvgLoss_a(X_t, X_nt) //计算平均损失
11.    if δ_min == δ then
12.        break
13.    else
14.        δ = δ_min
15.    end if
16. end for
```

4 实验和结果分析

4.1 实验设置

本文采用的数据集为英文数据集 Stanford Sentiment Treebank(SST-2)[17]和 AG news(ag_news)[18]、中文数据集 weibo 和 sogou[19]，这 4 个数据集是中英文领域公开且常用的数据集，具体数据集信息如表 1 所列。攻击的目标模型选择 TextCNN[20]和 BiLSTM[21]，英文采用 300 维的 Glove 词嵌入向量[22]，中文采用 300 维的 word2vec 词嵌入向量[23]。所有实验都是在训练集上搜索 CD-UAA 的扰动序列，并在测试集上评估其攻击效果。通过重复“the”(英文)、“的”(中文)来初始化扰动序列，扰动序列的长度为 3，每个批次的大小为 32，批次数量为 20。由于“@@PADDING@@”在词汇表中表示填充标记，无任何实际意义，因此当新的扰动序列中包含该标记时直接跳过；同时当分类任务为情感分析时，利用情感词库跳过情感含义明显的词汇。

表 1 实验数据集

Table 1 Experimental datasets

Datasets	Task	Language	Number of Classes	Train Data	Test Data	Average Length	TextCNN Accuracy /%	BiLSTM Accuracy /%
SST-2	sentiment Analysis	english	2	6920	1821	19.28	83.53	88.63
ag_news	news classification	english	4	120000	7600	43.64	92.76	93.87
weibo	sentiment Analysis	chinese	8	17000	4300	17.35	71.91	87.12
sogou	news classification	chinese	12	60000	10008	193.16	94.65	95.57

对抗攻击通常使用攻击成功率来衡量攻击效果。攻击成功率指目标模型对于对抗样本的分类错误率，这里用S_t和S_{nt}分别表示在X_t和X_{nt}上的攻击成功率，其中S_t越大越好，S_{nt}越小越好。

4.2 定性实验

在所有的实验中，将生成的扰动序列都添加到文本样本的前面。表 2 列出了在 BiLSTM 上攻击中英文情感分析的示例，即英文 SST-2(positive)、中文 weibo(难过，愤怒)，其中下划线为根据目标类别而生成的扰动序列，其余部分为原始文本，加粗部分为非目标类别的文本示例。如对于数据集 weibo，根据目标类别“难过，愤怒”生成的扰动序列为“订背倚逆差”，将其添加到文本样本的前面后，只有当 BiLSTM 对于原始文本的预测类别为“难过，愤怒”时，BiLSTM 的输出才会被混淆；预测类别为“开心，喜爱”等非目标类别时，扰动序列的影响会很小，甚至没有。

表 2 在 BiLSTM 上攻击 SST-2 和 weibo 的 CD-UAA 示例

Table 2 Examples of CD-UAA for attacking SST-2 and weibo on BiLSTM

Targeted Classes	Input(Underline=perturbation sequences)	Model Prediction
positive	cutting-room moldy untalented With a cast that includes some of the top actors working in independent film, Lovely& Amazing involves us because it is so incisive, so bleakly amusing about how we go about our lives.	positive(0.99)negative(0.97)
	cutting-room moldy untalented As a revenge thriller, the movie is serviceable, but it doesn't really deliver the delicious guilty pleasure of the better film versions.	**negative(0.99)negative(0.99)**
难过,愤怒	订背倚逆差明天早上4点去接老妈然后去长沙,感觉明天又会很累了,好多事情没有做。	难过(0.94)喜爱(0.97)
	订背倚逆差车位在本周五正式发售,结果大家的车昨天夜里就被砸了!	愤怒(0.99)恶心(0.99)
	订背倚逆差老师刚才让全体男生站起来集体说"祝各位公主节日快乐 越长越漂亮",笑死我了,哈哈哈,男生一张张苦逼脸太好笑了。	**开心(0.95)开心(0.94)**
	订背倚逆差那样我们就可以不再对着电话诉说想念,就可以每天清晨起来看见你的睡脸,然后一起吃一顿不太丰盛却很温暖的早餐。	**喜爱(0.99)喜爱(0.99)**

4.3 定量实验

将 Wallace 等[9]的 UAA 作为对比基线,在二分类任务中,探讨只攻击一个类别时 CD-UAA 与 UAA 的区别。将 SST-2 数据集中的 positive 和 negative 分别作为目标类别,使用相同的实验设置,实验结果如表 3 所列。

表 3 攻击 SST-2 单个类别的实验结果

Table 3 Experimental results of attacking single class on SST-2

(单位:%)

Model	Targeted Classes	UAA[9]		CD-UAA	
		S_t	S_{nt}	S_t	S_{nt}
TextCNN	positive	100	0	98.68	0
	negative	95.53	0	91.59	0
BiLSTM	positive	96.42	0	98.39	0
	negative	85.71	0	98.14	0

需要注意的是,UAA 的目标攻击指将其他类别的文本预测成目标类别,而 CD-UAA 指将目标类别的文本预测成其他类别,两者所描述的目标类别是相反的。从表 3 可以看出,CD-UAA 在 TextCNN 上对目标类别的攻击成功率 S_t 比 UAA 低 1.32%(positive)和 3.94%(negative),而在 BiLSTM 上比 UAA 高 1.97%(positive)和 12.43%(negative),同时两者对非目标类别的攻击成功率 S_{nt} 都为 0。

在攻击所有类别时,不存在非目标类别的文本,只有损失函数 $\mathcal{L}_t$ 会影响攻击效果,在多个数据集上与 UAA 进行对比,实验结果如表 4 所列,表中指标为 S_t。在新闻分类数据集 ag_news 和 sogou 上,使用不同损失函数 $\mathcal{L}_t$ 生成的所有扰动序列的攻击成功率都比 UAA 低;相反,在情感分析的数据集 SST-2 和 weibo 上,总有一个扰动序列的攻击成功率比 UAA 高,尤其是数据集 weibo。

表 4 攻击 4 个数据集所有类别的实验结果

Table 4 Experimental results of attacking all classes on four datasets

(单位:%)

	Algorithm	SST-2	weibo	ag_news	sogou
TextCNN	UAA[9]	49.97	16.14	75.66	57.80
	CD-UAA ($\mathcal{L}_t^{CE}$)	49.97	16.14	75.36	39.63
	CD-UAA ($\mathcal{L}_t^{LC}$)	49.84	84.51	74.81	9.53
	CD-UAA ($\mathcal{L}_t^{BL}$)	42.21	80.98	48.91	39.63
BiLSTM	UAA[9]	48.51	88.07	65.32	31.96
	CD-UAA ($\mathcal{L}_t^{CE}$)	49.26	92.85	38.38	31.96
	CD-UAA ($\mathcal{L}_t^{LC}$)	34.08	90.47	46.80	31.78
	CD-UAA ($\mathcal{L}_t^{BL}$)	46.65	92.63	55.82	30.49

在 ag_news 数据集上针对不同的目标类别,与 Wallace 等[9]的 UAA 进行对比,其中 UAA 只使用目标类别的文本,实验结果如表 5 所列,其中加粗部分表示 CD-UAA 对 ag_news 的整体攻击效果。对于 UAA 来说,即使只选择目标类别的文本,生成的扰动序列也会对非目标类别的文本产生很大的影响,其 S_{nt} 平均为 57.91%(TextCNN)和 40.84%(BiLSTM)。而通过使用分离的损失函数,其 S_{nt} 降低到 10%以下,同时对于目标类别的文本保持了较高的攻击成功率,S_t 的平均值比 UAA 高 4.46%(TextCNN)和 11.22%(BiLSTM)。总的来说,TextCNN 比 BiLSTM 有更大的 S_t 和更小的 S_{nt},更容易受到 CD-UAA 的攻击。

表 5 攻击 ag_news 不同目标类别的实验结果

Table 5 Experimental results of attacking different targeted classes on ag_news

(单位:%)

	Targeted Classes	UAA[9]		CD-UAA	
		S_t	S_{nt}	S_t	S_{nt}
TextCNN	Business	100	64.86	97.51	7.16
	Sci/Tech	99.83	60.5	98.09	1.84
	Sports	100	66	100	6.47
	World	100	65.61	100	7.24
	Business, Sci/Tech	100	8.28	99.06	3.08
	Business, Sci/Tech, Sports	67.88	82.19	99.81	0
	Avg	**94.62**	**57.91**	**99.08**	**4.30**
BiLSTM	Business	95.04	44.26	91.09	7.96
	Sci/Tech	22.25	18.22	69.85	8.50
	Sports	70.41	52.71	52.81	27.98
	World	80.06	51.16	70.11	3.36
	Business, Sci/Tech	28.72	28.37	49.58	9.19
	Business, Sci/Tech, Sports	18.88	50.33	49.21	0
	Avg	**52.56**	**40.84**	**63.78**	**9.50**

在中文数据集 weibo 和 sogou 上攻击不同的目标类别,观察 CD-UAA 的攻击效果,实验结果如表 6 和表 7 所列,其中加粗部分表示 CD-UAA 对 weibo 和 sogou 的整体攻击效果。在 weibo 上,两个模型的 S_t 都保持在 82%以上;而 S_{nt} 平均为 8.57%(TextCNN)和 20.48%(BiLSTM)。在 sogou 上,S_t 平均为 68.42%(TextCNN)和 43.12%(BiLSTM);S_{nt} 平均为 19.31%(TextCNN)和 19.02%(BiLSTM)。总的来说,weibo 更容易受到 CD-UAA 的攻击。

表 6 攻击 weibo 不同目标类别的实验结果

Table 6 Experimental results of attacking different targeted classes on weibo

Targeted Classes	TextCNN			BiLSTM		
	Perturbation Sequences	S_t/%	S_{nt}/%	Perturbation Sequences	S_t/%	S_{nt}/%
惊讶	公积金 38 节 的	100	9.63	练习曲 祝 棕子	88.89	16.93
喜爱	坏账 揍 可不带	100	7.93	苦下 允浩俊秀][	92.86	24.91
开心	北约 厚外 沃尔玛 25 万	93.59	8.43	怂掉 艺员 2001 年	86.46	16.32
愤怒	淡淡 太 out 朵	100	8.34	帅点儿 圣训 执业	97.24	19.10
惊讶,无	真是 !!!! 冰淇林	97.23	17.44	童子 试问 尼玛	85.92	35.47
喜爱,开心	被告人 内伤 定价	100	7.91	碰对 就算是 碰对	93.62	21.41
恶心,害怕	4414 亿 胡耀邦 的	96.30	4.12	交界 平平淡淡才是真 宪(	82.85	14.83
难过,愤怒	圆融 ~[反对党	95.81	4.76	订 背倚 逆差	91.82	14.88
Avg	—	**97.87**	**8.57**	—	**89.96**	**20.48**

表 7 攻击 sogou 不同目标类别的实验结果

Table 7 Experimental results of attacking different targeted classes on sogou

Targeted Classes	TextCNN			BiLSTM		
	Perturbation Sequences	S_t/%	S_{nt}/%	Perturbation Sequences	S_t/%	S_{nt}/%
women	通辽 汽油机 财税	88.98	19.09	搜狐 COOL 车型	35.96	25.26
news	托吉安岛 OBS 蛋不容易	99.72	7.35	搜狐 m²〔 m²〔	79.55	22.79
auto	早樱 58236831 汪泽	54.63	15.35	搜狐 搜狐 romp	45.15	25.10
sports	童装 悬架 位价	77.15	20.58	搜狐 COOL 戛纳	59.57	11.61
women,house,it	达斡尔 车型 江大红	45.80	27.41	责任 津报网 真题	8.07	7.14
yule,health,news	搜狐 轴距 8/37	79.61	28.88	搜狐 m²〔 COOL	47.14	26.75
cul,business,travel	杭德罗·普林斯 报既 出铜	51.81	10.34	责任 责任 责任	21.25	5.85
learning,auto,sports	林青霞 早樱 林青霞	49.67	25.49	搜狐 COOL 马莹莹	48.26	27.68
Avg	—	**68.42**	**19.31**	—	**43.12**	**19.02**

4.4 消融实验

4.4.1 损失函数 $\mathscr{L}_t$ 和 $\mathscr{L}_{nt}$

损失函数 $\mathscr{L}_t$ 和 $\mathscr{L}_{nt}$ 选择不同的形式会直接影响到扰动序列的攻击效果,将 weibo 的“无”和 ag_news 的“Business”分别作为攻击的目标类别,并选用不同的目标模型,实验结果如表 8 所列,其中每行的指标是 S_t 和 S_{nt},加粗部分表示 S_t 和 S_{nt} 的差值最大。相比只使用 $\mathscr{L}_t$,$\mathscr{L}_{nt}$ 可以有效减弱对非目标类别的影响,而 $\mathscr{L}_t^{CE}$ 相对于其他损失函数,对非目标类别的影响更大。总的来说,对于不同的目标类别以及不同的目标模型,其最佳的损失函数配置是不同的,因此找到一个通用的损失函数配置是比较困难的。为了解决该问题,可以遍历所有的损失函数配置,生成若干个扰动序列,利用生成扰动序列的训练数据子集,从中选择 S_t 和 S_{nt} 差值最大的扰动序列,这个操作对于攻击者是可选的,因为这会耗费更多的计算资源和时间。在后续的消融实验中,为了消除损失函数的影响,损失函数 $\mathscr{L}_t$ 选择 $\mathscr{L}_t^{BL}$,损失函数 $\mathscr{L}_{nt}$ 选择 $\mathscr{L}_{nt}^{CE}$。而在其他实验中,都采用该可选操作。

表 8 不同损失函数配置对 S_t 和 S_{nt} 的影响

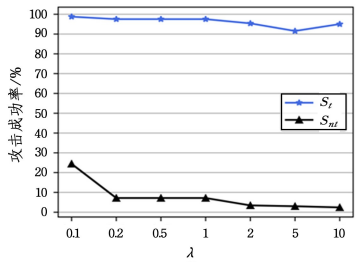
Table 8 Influence of different loss function configurations on S_t and S_{nt}

(单位:%)

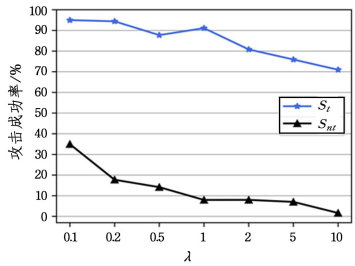
	$\mathscr{L}_{nt}$	$\mathscr{L}_t$					
		$\mathscr{L}_t^{CE}$		$\mathscr{L}_t^{LC}$		$\mathscr{L}_t^{BL}$	
TextCNN+ ag_news	—	100	67.81	100	40.49	100	67.81
	$\mathscr{L}_{nt}^{CE}$	100	64.42	100	21.24	**97.51**	**7.16**
	$\mathscr{L}_{nt}^{LC}$	100	65.01	100	59.68	91.11	2.20
	$\mathscr{L}_{nt}^{BL}$	100	27.77	99.64	31.82	90.94	9.38
BiLSTM+ weibo	—	99.29	79.69	97.91	64.57	97.83	75.14
	$\mathscr{L}_{nt}^{CE}$	96.57	53.59	91.87	45.5	83.27	42.28
	$\mathscr{L}_{nt}^{LC}$	99.29	79.69	95.66	62.26	84.54	43.85
	$\mathscr{L}_{nt}^{BL}$	**79.68**	**22.13**	**84.14**	**26.59**	75.62	34.76

4.4.2 权重参数 λ

对于 ag_news 的类别“Business”,在两个模型上评估式(3)中的 λ 对 S_t 和 S_{nt} 的影响,结果如图 2 所示。随着 λ 的增大,S_{nt} 在逐渐减小,这意味着生成的扰动序列更加隐蔽。由于 S_t 也会随之减小,因此并不能带来更好的攻击效果,尤其是 BiLSTM。



(a) TextCNN



(b) BiLSTM

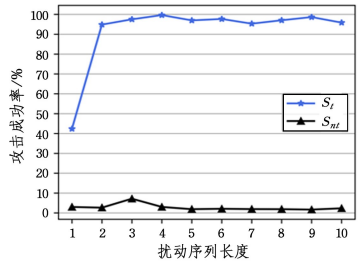
图 2 式(3)中的 λ 对 S_t 和 S_{nt} 的影响

Fig. 2 Influence of λ in Eq. 3 on S_t and S_{nt}

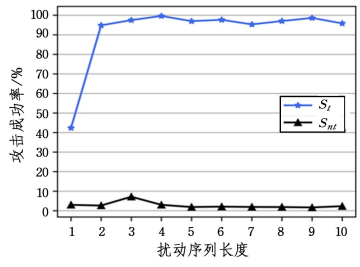
4.4.3 扰动序列长度

扰动序列的长度不仅会影响人们对原始文本的理解,而且对 S_t 和 S_{nt} 的影响也非常大。对于 ag_news 的“Business”类别,两个模型的结果如图 3 所示。序列长度为 2 或者 3 时,对目标类别的攻击成功率 S_t 急剧上升,然后随着序列长度的

增加，S_t 一直保持在 90%以上，对非目标类别的攻击成功率 S_{nt} 保持在 10%以下。在实际的攻击过程中，可以利用该算法生成多个长短不一的扰动序列，根据原始文本的长短而选择合适长度的扰动序列，这样可以在保证攻击效果的同时，降低对原始文本的影响。



(a)TextCNN



(b)BiLSTM

图 3 扰动序列长度对 S_t 和 S_{nt} 的影响

Fig. 3 Influence of perturbation sequences length on S_t and S_{nt}

4.4.4 批次数量

在 CD-UAA 中，使用不同的训练数据会生成不同的扰动序列，而训练数据的批次数量会控制训练数据的量，进而影响攻击效果。对于 ag_news 的“Business”类别，在两个模型上评估批次数量对 S_t 和 S_{nt} 的影响，结果如表 9 所列，其中每个批次的大小为 32。从表 9 可以看出，CD-UAA 使用 32 条训练数据就可以达到较好的攻击效果，这足以说明算法的高效性。如在 TextCNN 上，1 个批次的 S_t 为 91.65%，只比 5 个批次的 97.51%低了 5.86%；S_{nt} 为 6.15%，比 5 个批次的 7.16%更低。

表 9 批次数量对 S_t 和 S_{nt} 的影响

Table 9 Influence of number of batch on S_t and S_{nt}

(单位：%)

Number of batch	TextCNN		BiLSTM	
	S_t	S_{nt}	S_t	S_{nt}
1	91.65	6.15	80.81	16.31
5	97.51	7.16	91.68	10.99
10	97.51	7.16	78.69	11.03
20	97.51	7.16	91.09	7.96
40	94.31	4.74	87.01	6.65

4.5 迁移实验

观察 CD-UAA 的扰动序列在两个模型之间的迁移性，即选定攻击的目标类别，在原模型上生成扰动序列，随后观察该序列在目标模型上的攻击效果，结果如表 10 所列，其中加粗部分表示扰动序列在目标模型上 S_t 和 S_{nt} 的差值比原模型大。扰动序列具有一定的可迁移性，在 BiLSTM 上生成的扰动序列可以很好地迁移到 TextCNN 中，反之则效果很差，尤其对于数据集 weibo 来说，扰动序列从 BiLSTM 迁移到 TextCNN，S_t 可以达到 82.5%，S_{nt} 为 6.35%。

结束语 针对目前通用对抗攻击的局限性，本文提出了一种面向文本分类的类别区分式通用对抗攻击方法(CD-UAA)，可以对目标类别的文本样本达到较高的攻击成功率，同时对非目标类别的文本样本有较低的攻击成功率。通过在多个中英文数据集上的大量实验，验证了该攻击方法的有效性。由于未对生成的扰动序列做任何语义上的搜索限制，导致扰动序列在语义上并不流畅，未来可以研究如何找到流畅语义且攻击成功的扰动序列。同时，对于 CD-UAA，目前还缺乏检测和防御的相关方法。

表 10 CD-UAA 的迁移性

Table10 Transferability of CD-UAA

Targeted Classes	Original Model	Perturbation Sequences	S_t/%	S_{nt}/%	Targeted Model	S_t/%	S_{nt}/%
ag_news (Business, Sports)	TextCNN	cybersecurity parchin seabird	98.15	1.69	BiLSTM	10.52	8.65
	BiLSTM	washingtonpost. com mydoom mydoom	82.71	21.23	TextCNN	**95.51**	**31.64**
weibo (喜爱，难过)	TextCNN	审理 公款 [哈哈]	100	5.01	BiLSTM	34.76	13.83
	BiLSTM	报价 碰对 树枝	95.33	20.90	TextCNN	**82.50**	**6.35**
sogou (women, yule)	TextCNN	通辽 变速箱 财税	93.80	28.04	BiLSTM	3.35	1.70
	BiLSTM	搜狐 m²〔 车型	31.91	26.60	TextCNN	**29.85**	**15.41**

参 考 文 献

[1] 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. Stockholm: AAAI Press, 2018: 4208-4215.

[2] WANG W Q, WANG R, WANG L N, et al. Adversarial examples generation approach for tendency classification on Chinese texts[J]. Ruan Jian Xue Bao/Journal of Software, 2019, 30(8): 2415-2427.

[3] TONG X, WANG L N, WANG R Z, et al. A Generation Method of Word-level Adversarial Samples for Chinese Text Classification[J]. Netinfo Security, 2020, 20(9): 12-16.

[4] CHENG M, YI J, CHEN P Y, et al. Seq2sick: Evaluating the robustness of sequence-to-sequence models with adversarial examples[C]// Proceedings of the AAAI Conference on Artificial Intelligence. New York: AAAI Press, 2020, 34(4): 3601-3608.

[5] ILYAS A, SANTURKAR S, TSIPRAS D, et al. Adversarial examples are not bugs, they are features[C]// Proceedings of the 33rd International Conference on Neural Information Processing

Systems. Red Hook:Curran Associates Inc,2019:125-136.

[6] ZHANG C,BENZ P,IMTIAZ T,et al. Understanding adversarial examples from the mutual influence of images and perturbations[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle:IEEE Press,2020:14521-14530.

[7] TONG X,WANG B J,WANG R Z,et al. Survey on Adversarial Sample of Deep Learning Towards Natural Language Processing [J]. Computer Science,2021,48(1):258-267.

[8] BEHJATI M,MOOSAVI-DEZFOOLI S M,BAGHSHAH M S,et al. Universal adversarial attacks on text classifiers[C]// 2019 IEEE International Conference on Acoustics,Speech and Signal Processing(ICASSP 2019). Brighton:IEEE Press,2019:7345-7349.

[9] WALLACE E,FENG S,KANDPAL N,et al. Universal Adversarial Triggers for Attacking and Analyzing NLP[C]// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong:ACL Press,2019:2153-2162.

[10] SONG L,YU X,PENG H T,et al. Universal Adversarial Attacks with Natural Triggers for Text Classification[C]// Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics:Human Language Technologies. Online:ACL Press,2021:3724-3733.

[11] MOOSAVI-DEZFOOLI S M,FAWZI A,FAWZI O,et al. Universal adversarial perturbations[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York:IEEE Press,2017:1765-1773.

[12] HEIDENREICH H S,WILLIAMS J R. The Earth Is Flat and the Sun Is Not a Star:The Susceptibility of GPT-2 to Universal Adversarial Triggers[C]// Proceedings of the 2021 AAAI/ACM Conference on AI,Ethics,and Society. New York:ACM Press,2021:566-573.

[13] GUPTA T,SINHA A,KUMARI N,et al. A method for computing class-wise universal adversarial perturbations[J]. arXiv:1912.00466,2019.

[14] ZHANG C,BENZ P,IMTIAZ T,et al. Cd-uap:Class discriminative universal adversarial perturbation[C]// Proceedings of the AAAI Conference on Artificial Intelligence. New York:AAAI Press,2020,34(4):6754-6761.

[15] BENZ P,ZHANG C,IMTIAZ T,et al. Double targeted universal adversarial perturbations[C]// Proceedings of the Asian Conference on Computer Vision. Kyoto:ACCVPress,2020:1-17.

[16] EBRAHIMI J,RAO A,LOWD D,et al. HotFlip:White-Box Adversarial Examples for Text Classification[C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2:Short Papers). Melbourne:ACL Press,2018:31-36.

[17] SOCHER R,PERELYGIN A,WU J,et al. Recursive deep models for semantic compositionality over a sentiment treebank [C]// Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Seattle:ACL Press,2013:1631-1642.

[18] ZHANG X,ZHAO J,LECUN Y. Character-level convolutional networks for text classification[C]// Proceedings of the 28th International Conference on Neural Information Processing Systems (Volume 1). Cambridge,MA,USA:MIT Press,2015:649-657.

[19] LI L,SHAO Y,SONG D,et al. Generating Adversarial Examples in Chinese Texts Using Sentence-Pieces[J]. arXiv:2012.14769,2020.

[20] KIM Y. Convolutional neural networks for sentence classification[C]// Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha:ACL Press,2014.1746-1751.

[21] MCCANN B,BRADBURY J,XIONG C,et al. Learned in Translation:Contextualized Word Vectors[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook,NY,USA:Curran Associates Inc,2017:6297-6308.

[22] PENNINGTON J,SOCHER R,MANNING C D. Glove:Global vectors for word representation[C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing(EMNLP). Doha:ACL Press,2014:1532-1543.

[23] LI S,ZHAO Z,HU R,et al. Analogical Reasoning on Chinese Morphological and Semantic Relations[C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Short Papers). Melbourne:ACL Press,2018:138-143.

HAO Zhi-rong, born in 1997, postgraduate. His main research interests include adversarial examples and natural language processing.

CHEN Long, born in 1970, professor, Ph.D supervisor. His main research interests include digital forensics and AI security.

(责任编辑:喻藜)