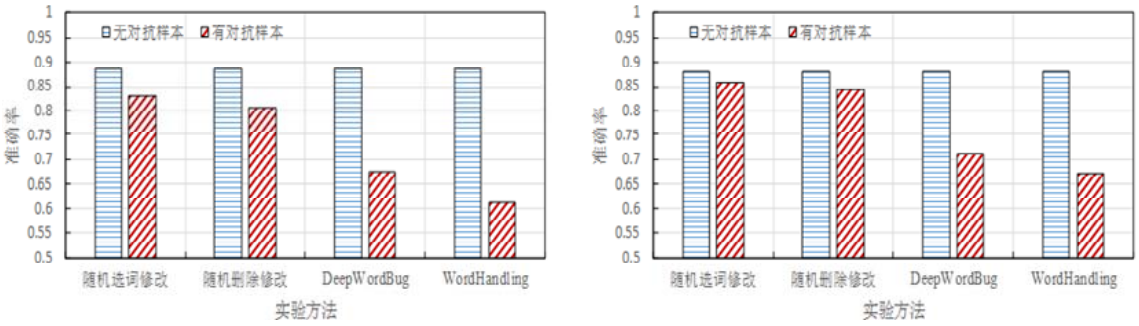


(a) LSTM 模型对比实验 (b) CNN 模型对比实验

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

图 6 表 2 中京东数据上对抗样本检测结果

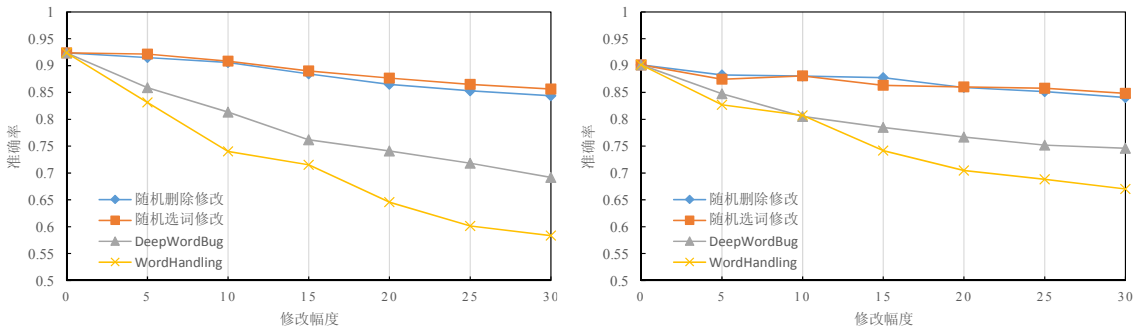
由表 2 可以看出,关键词语对输入数据类别倾向的影响较大,使用随机的方式对输入进行改动得到的结果并不理想;而且本文提出的 WordHandling 算法比 DeepWordBug 效果更佳.表 3 则是选取的若干原始样本和在其基础之上生成的对抗样本的例子,由表 3 可以看出,生成的对抗样本仍然能够通过语义上下文被人所理解,文本意思变化在可接受范围内.

Table 3 Examples of original examples and generated adversarial examples

表 3 原始样本和生成的对抗样本例子

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

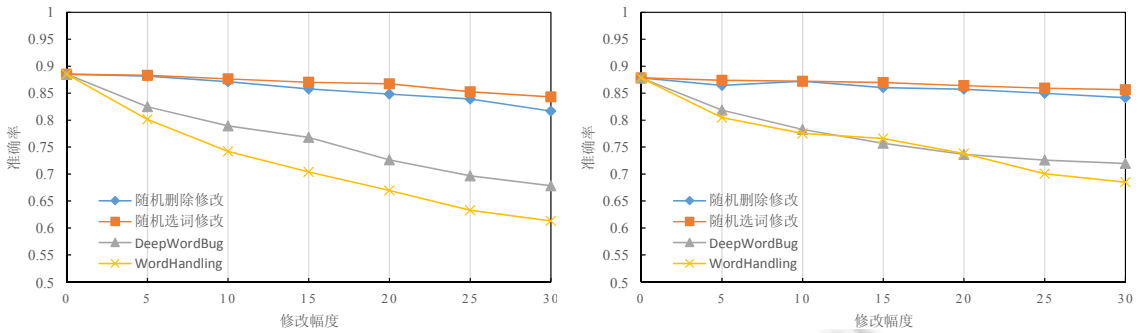
为了验证模型对于对抗样本检测的准确率与样本修改的幅度 m 的关联性,从两种数据集中分别选取 1 000 条长度大于 120 字的数据,根据不同的修改幅度生成相应的对抗样本.图 7 为携程酒店评论数据集在两种模型上检测的准确率随修改规模 m 变化的曲线,图 8 则是在京东购物评论数据集在两种模型上的实验结果.



(a) LSTM 模型对比实验 (b) CNN 模型对比实验

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

图 7 携程数据对抗样本检测准确率随修改幅度 m 的变化曲线



(a) LSTM 模型对比实验 (b) CNN 模型对比实验

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

图 8 京东数据抗样本检测准确率随修改幅度 m 的变化曲线

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

4.3 对抗样本质量度量

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

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

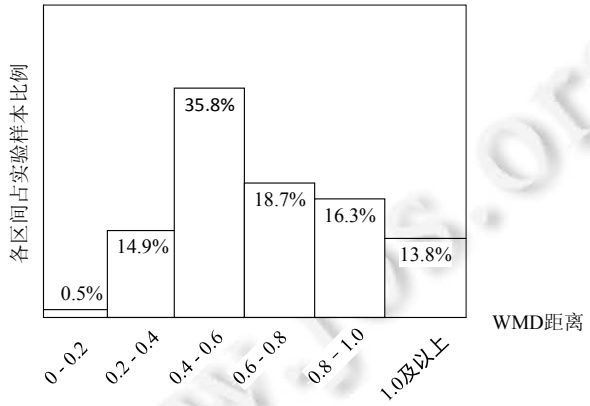


Fig.9 Ratio of sample numbers to total samples in different WMD distance intervals

图 9 不同 WMD 距离区间内样本数量占总样本的比例

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

例).这也导致与相同修改幅度的长文本相比,短文本的可读性稍差.

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

表 4 不同 WMD 距离区间长短文本数量所占该区间比例

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

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

Table 5 Examples of WMD distance calculation

表 5 WMD 距离计算实例

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

5 总 结

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

References:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

附中文参考文献:

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

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

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

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

汪润(1991－),男,博士,主要研究领域为移动设备隐私保护,机器学习.

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