文本对抗攻击入坑宝典

原创 阿毅 夕小瑶的卖萌屋 2021-02-26 17:00

文 | 阿毅
编 | 小轶

如果是咱家公众号的忠实粉丝就一定还记得之前咱家一篇关于**NLP Privacy的文章**，不出意外的话，你们是不是现在依然还担心自己的隐私被输入法窃取而瑟瑟发抖。所以，我们又来了！今天给大家讨论的是NLP Privacy中一个非常核心的话题——文本对抗攻击。

相信大家已经非常熟悉**对抗攻击**了，此类攻击是攻击者针对机器学习模型的输入即数值型向量（*Numeric Vectors*）设计的一种可以让模型做出误判的攻击。简言之，对抗攻击就是生成对抗样本的过程。**对抗样本**的概念最初是在2014年提出的，指的是一类人为构造的样本，通过对原始的样本数据添加针对性的微小扰动所得到（该微扰不会影响人类的感知），但会使机器学习模型产生错误的输出[1]。因此，从上述定义可知，对抗攻击以及对抗样本的生成研究最开始被用于计算机视觉领域。在当时，那家伙，文章多的你看都看不完...当然在这里我也抛出当时写的比较好的一篇综述：**“Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey”**[2]。大家可以温故而知新啦。

当视觉领域中的对抗攻击研究很难再有重大突破的时候（坑已满，请换坑），研究人员便把目光转移到了NLP领域。其实就NLP领域而言，垃圾邮件检测、有害文本检测、恶意软件查杀等实用系统已经大规模部署了深度学习模型，安全性对于这些系统尤为重要。但相比于图像领域，NLP领域对抗攻击的研究还远远不够，特别是文本具有离散和前后输入具有逻辑的特点使得对抗样本的生成更具挑战性，也有更多的研究空间。我们欣喜地看到，目前有越来越多的 NLP 研究者开始探索文本对抗攻击这一方向，以 2020 年 ACL 为例，粗略统计有超过 10 篇相关论文，其中最佳论文 **Beyond Accuracy: Behavioral Testing of NLP Models with CheckList** [3]中大部分测试方法其实和文本对抗攻击有

异曲同工之妙。故在本次推文中，我们一起来探究和领略一下如何在NLP领域实施对抗攻击，并提供一些在该领域继续深入挖掘的工具和方向。

对抗攻击的分类

对抗攻击按攻击者所掌握的知识来分的话，可分为以下两类：

- **白盒攻击**：称为*white-box attack*，也称为*open-box attack*，即攻击者对模型（包括参数、梯度等信息）和训练集完全了解，这种情况比较攻击成功，但是在实际情况中很难进行操作和实现。
- **黑盒攻击**：称为*black-box attack*，即攻击者对模型不了解，对训练集不了解或了解很少。这种情况攻击很难成功但是与实际情况比较符合，因此也是主要的研究方向。

如果按攻击者的攻击目标来分的话，可以分为以下两类：

- **定向攻击**：称为*targeted attack*，即对于一个多分类网络，把输入分类误判到一个指定的类上
- **非定向攻击**：称为*non-target attack*，即只需要生成对抗样本来欺骗神经网络，可以看作是上面的一种特例。

发展历史与方法分类

我们先谈谈白盒攻击，因为白盒攻击易于实现，因此早在2014年关于对抗样本的开山之作“**Intriguing Properties of Neural Networks**”中设计了一种基于梯度的白盒攻击方法。具体来说，作者通过寻找最小的损失函数添加项，使得神经网络做出误分类，将问题转化成了凸优化。问题的数学表述如下：

$$\begin{aligned}&\min ||r||_2:\\&1.f(x+r)=l\\&2.x+r\in[0,1]^m\end{aligned}$$

$f(x)$表示习得的分类映射函数，r表示改变的步长，公式表达了寻找使得$f(x+r)$映射到指定的类l上的最小的r。在此之后，许多研究人员在上述方法的基础上提出了许多改进的基于梯度的方法，具体可见

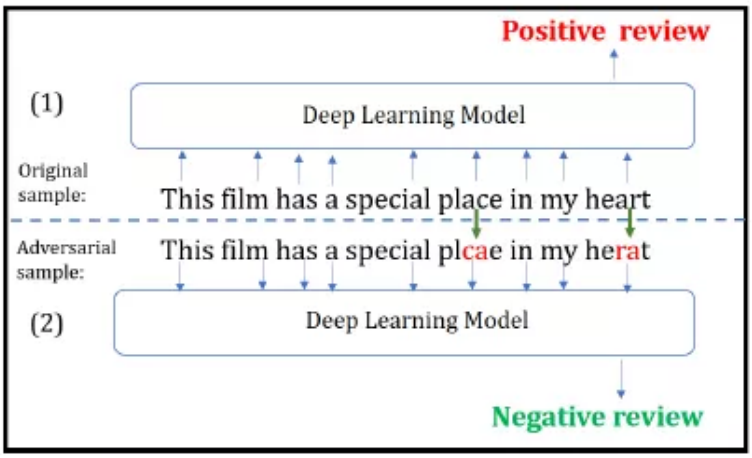
[4-6]。

> 后来，研究人员逐渐从白盒攻击的研究转向研究黑盒攻击，*Transfer-based*方法就是过渡时期的产物。Nicolas Papernot等人在2017年的时候利用训练数据可以训练出从中生成对抗性扰动的完全可观察的替代模型[7]。因此，基于Transfer的攻击不依赖模型信息，但需要有关训练数据的信息。**此外，[8]文献证明了如果在一组替代模型上生成对抗性样本，则在某些情况下，模型被攻击的成功率可以达到100%（好家伙，100%真厉害）**。近几年，不同类型的攻击方法越来越多，但总体来说归为以下三类：Score-based方法、Decision-based方法、**Attack on Attention方法[9]**（这个方法非常新，有坑可跳），前两大类方法的相关研究和参考文献可阅读原文一探究竟，在这里不再赘述。

文本对抗攻击

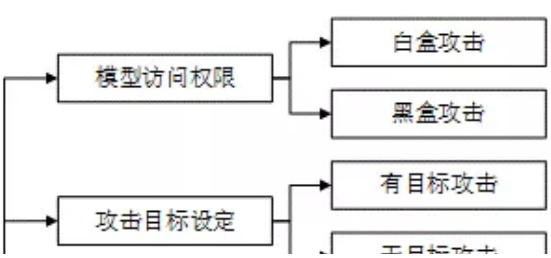
基本概念

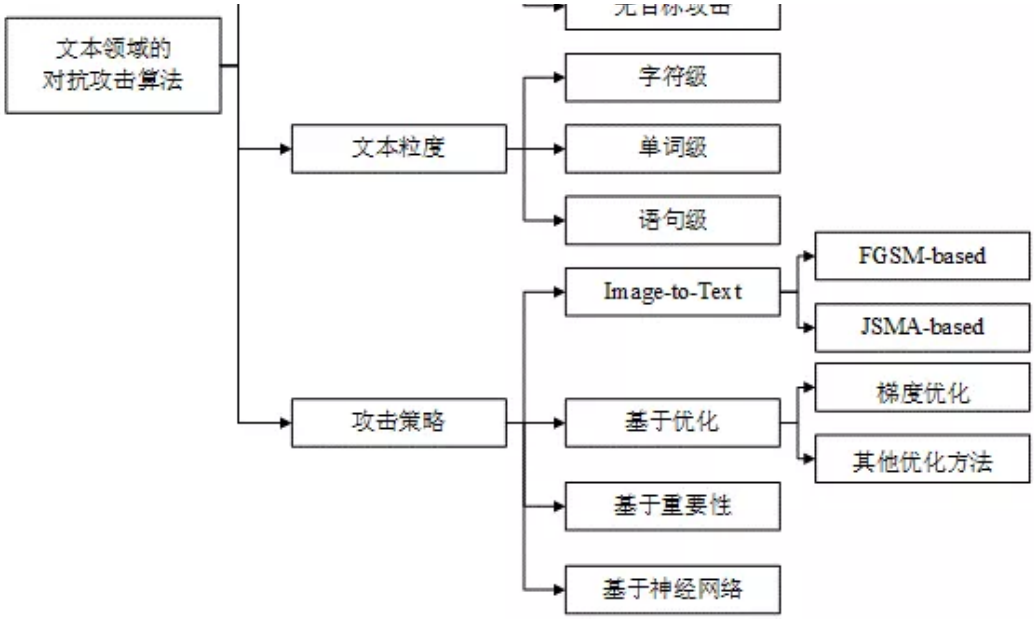
下图展示了文本领域内实现对抗攻击的一个例子。语句（1）为原始样本，语句（2）为经过几个字符变换后得到的对抗样本。深度学习模型能正确地将原始样本判为正面评论，而将对抗样本误判为负面评论。而显然，这种微小扰动并不会影响人类的判断。



算法的分类

首先，根据上述对抗攻击的分类。同样地，文本中的对抗攻击也可以分为**黑盒攻击和白盒攻击**。除此之外，由于文本涉及到字符、词汇、句子。因此我们可以根据添加扰动时所操作的文本粒度可以分为**字符级、单词级和语句级**攻击。具体来说，字符级攻击是通过插入、删除或替换字符，以及交换字符顺序实现；单词级攻击主要通过替换单词实现，基于近义词、形近词、错误拼写等建立候选词库；语句级攻击主要通过文本复述或插入句子实现。具体分类详见下图.





攻击方式的发展和分类

根据攻击策略和攻击方式我们可以分为*Image-to-Text*（借鉴图像领域的经典算法）、基于优化的攻击、基于重要性的攻击以及基于神经网络的攻击。*Image-to-Text*攻击方式的思想是将文本数据映射到连续空间，然后借鉴图像领域的一些经典算法如*FGSM*、*JSMA*等，生成对抗样本；基于优化的攻击则是将对抗攻击表述为带约束的优化问题，利用现有的优化技术求解，如梯度优化、**遗传算法优化**；基于重要性的攻击通常首先利用梯度或文本特性设计评分函数锁定关键词，然后通过文本编辑添加扰动；基于神经网络的攻击训练神经网络模型自动学习对抗样本的特征，从而实现对抗样本的自动化生成。具体的算法细节大家可移步一篇写的非常全面的综述“**Adversarial Attacks on Deep Learning Models in Natural Language Processing: A Survey**“。

文本对抗攻击相关资源

文献总结

如下图所示，清华大学自然语言处理与社会人文计算实验室（THUNLP）总结了各类文本对抗领域的相关文献，其中包含但不限于工具包、综述、文本对抗攻击、文本对抗防御、模型鲁棒性验证、基准和评估等内容。针对本文涉及的文本对抗攻击领域，该列表收录了句级、词级、字级、混合四个子部分，并且还为每篇论文打上了受害模型可见性的标签：

gradient/score/decision/blind

除了提供论文 pdf 链接之外，如果某篇论文有公开代码或数据，也会附上相应的链接[19]。

其中必须的综述论文如下：

> -- Analysis Methods in Neural Language Processing: A Survey. Yonatan Belinkov, James Glass. TACL 2019.
> -- Towards a Robust Deep Neural Network in Text Domain A Survey. Wenqi Wang, Lina

> Wang, Benxiao Tang, Run Wang, Aoshuang Ye. 2019.
> -- Adversarial Attacks on Deep Learning Models in Natural Language Processing: A Survey. Wei Emma Zhang, Quan Z. Sheng, Ahoud Alhazmi, Chenliang Li. 2019.

Must-read Papers on Textual Adversarial Attack and Defense (TAAD)

last commit november 2020 | PaperNumber 70 | PRs Welcome

Mainly Contributed and Maintained by Fanchao Qi, Chenghao Yang and Yuan Zang.

Great thanks to other contributors Di Jin, Boxin Wang and Jingkang Wang! (names are not listed in particular order)

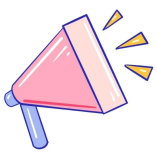
Contents

文本对抗攻击工具包

目前文本攻击工具包为该领域的研究人员提供了非常好的开发和研究基础。这里介绍两个比较常用的：

- 清华大学自然语言处理与社会人文计算实验室开源的*OpenAttack*[20]
- 弗吉尼亚大学祁妍军教授领导的 Qdata 实验室开发的*TextAttack*[21]

至于如何使用上述两种工具包，请大家火速前往项目主页一探究竟，并不要忘了给一个Star哦！！！



参考文献

[1]. Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks[J]. arXiv preprint arXiv:1312.6199, 2013.

[2]. Akhtar N, Mian A. Threat of adversarial attacks on deep learning in computer vision: A survey[J]. IEEE Access, 2018, 6: 14410-14430.

[3]. Ribeiro M T, Wu T, Guestrin C, et al. Beyond accuracy: Behavioral testing of NLP models with CheckList[J]. arXiv preprint arXiv:2005.04118, 2020.

[4]. Tramèr F, Kurakin A, Papernot N, et al. Ensemble adversarial training: Attacks and defenses[J]. arXiv preprint arXiv:1705.07204, 2017.

[5]. Moosavi-Dezfooli S M, Fawzi A, Frossard P. Deepfool: a simple and accurate method to fool deep neural networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2574-2582.

[6]. Papernot N, McDaniel P, Goodfellow I. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples[J]. arXiv preprint arXiv:1605.07277, 2016.

[7]. Papernot N, McDaniel P, Goodfellow I, et al. Practical black-box attacks against machine learning[C]//Proceedings of the 2017 ACM on Asia conference on computer and communications security. 2017: 506-519.

[8]. Lu J, Issaranon T, Forsyth D. Safetynet: Detecting and rejecting adversarial examples robustly[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 446-454.

[9]. Chen S, He Z, Sun C, et al. Universal adversarial attack on attention and the resulting dataset damagenet[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020.

[10]. https://www.secrss.com/articles/25644

[11]. 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 (TIST), 2020, 11(3): 1-41.

[12]. Cheng M, Le T, Chen P Y, et al. Query-efficient hard-label black-box attack: An optimization-based approach[J]. arXiv preprint arXiv:1807.04457, 2018.

[13]. Brendel W, Rauber J, Bethge M. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models[J]. arXiv preprint arXiv:1712.04248, 2017.

[14]. Mrkšić N, Séaghdha D O, Thomson B, et al. Counter-fitting word vectors to linguistic constraints[J]. arXiv preprint arXiv:1603.00892, 2016.

[15]. Alzantot M, Sharma Y, Elgohary A, et al. Generating natural language adversarial examples[J]. arXiv preprint arXiv:1804.07998, 2018.

[16]. https://www.secrss.com/articles/25644

[17]. https://www.jiqizhixin.com/articles/2019-06-10-6

[18]. https://www.aminer.cn/research_report/5f50600e3c99ce0ab7bcb539

[19]. https://github.com/thunlp/TAADpapers

[20]. http://nlp.csai.tsinghua.edu.cn/project/openattack/

[21]. https://github.com/QData/TextAttack