智源学者成果展示——自然语言处理

作者　臧　原、岂凡超、刘知远、孙茂松（清华大学）

2020年6月

项目负责人：臧原、岂凡超、刘知远、孙茂松

单位 / 团队：清华大学自然语言处理与社会人文计算实验室

成果名称：基于义原知识和离散粒子群优化算法的文本对抗攻击

成果简介：

对抗攻击（adversarial attack）旨在对目标模型的原始输入进行轻微扰动以生成对抗样本（adversarial example），进而使目标模型判断出错。目前大量实验已经证明深度学习模型易受对抗攻击的影响，例如对恶意评论进行些许无关紧要的修改就可以骗过谷歌的恶评检测系统。此外，对抗攻击的研究也有利于反过来提高模型的鲁棒性和可解释性。

在自然语言处理领域，随着深度学习模型在诸如垃圾邮件过滤等实用系统中的大规模部署，其安全性和鲁棒性也越来越重要。相比于图像、声音等领域，由于文本的离散特性，文本领域的对抗攻击更具挑战性。任何对文本的轻微扰动，即使小到一个字符，都有可能破坏文本的语义和语法性，甚至改变其真实的分类标签而使得攻击无效。

现有的文本对抗攻击方法可以根据其产生的扰动类型分为句级、词级和字符级。其中词级对抗攻击方法在攻击成功率、攻击有效性、对抗样本质量等方面有相对更好的整体性能。我们提出，词级文本对抗本质上是一个两步的组合优化问题：（1）第一步为搜索空间的构建，即确定原始输入中每个词的候选替换词集并将其组合形成一个离散的搜索空间；（2）第二步为在上述离散空间中搜索对抗样本。

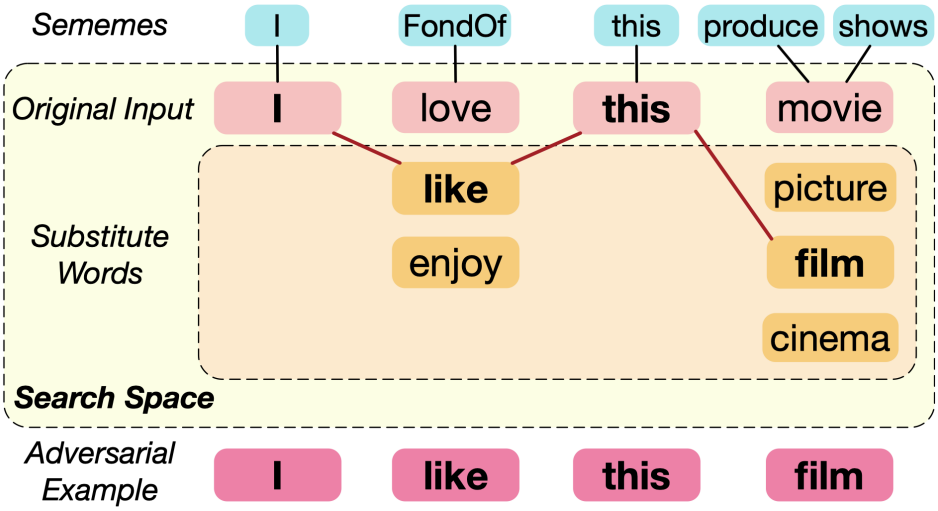

图 1　作为组合优化问题的词级文本对抗攻击。原始输入为 I love this movie，首先为每个词找到若干合适的候选替换词（例如，为 love 找到了 like 和 enjoy），然后在每个词本身及其候选替换词的所有组合构成的离散空间中搜索能够成功攻击目标模型的对抗样本。

我们在这两步上均提出了新的方法。对于第一步，我们提出了基于义原的词替换策略，相比于已有的基于词向量、同义词等词替换方法，其能够对更多的词找到更加丰富而合适的替换词，从而构建更大的搜索空间。对于第二步，我们提出了基于离散粒子群优化（particle swarm optimization）的对抗样本搜索算法，比已有的基于贪心、遗传算法等搜索方法具有更高的搜索效率和更好的搜索效果。

我们在情感分析、自然语言推断等任务上对 BiLSTM 和 BERT 这两个使用最广泛的模型进行攻击，实验结果表明无论是在攻击成功率还是在对抗样本质量方面，我们的模型都显著优于现有的其他模型，其中在 IMDB 这个情感分析数据集上对 BiLSTM 的攻击成功率达到了 100%。

此外我们也通过实验证明了基于义原的词替换策略以及基于离散粒子群优化的搜索算法的优越性。例如，基于义原的词替换策略平均为每个词找到 10–13 个候选替换词，而基于词向量或同义词的词替换方法只能找到 3–4 个候选替换词，而且基于义原的词替换策略找到的候选替换词的质量也更高。

She breaks the pie dish and screams out that she is not handicapped.		
Embedding/LM	Synonym	Sememe
tart, pizza, apple, shoemaker, cake cheesecake	None	cheese, popcorn, ham, cream, break, cake, pizza, chocolate, and 55 more

图 2　基于义原的词替换策略为 pie 这个词找到了 60 多个合适的候选替换词，基于词向量的方法只找到 5 个，而基于同义词的方法未能找到任何替换词。

我们还发现我们的攻击模型产生的对抗样本有更高的迁移性，即使用对目标模型 A 有效的对抗样本来攻击另一个目标模型 B 有更高的成功率。最后，我们还在对抗训练的实验中，验证了我们的模型产生的对抗样本加入到训练数据中能够给目标模型带来更多的鲁棒性的提升。

论文链接：https://arxiv.org/pdf/1910.12196.pdf

Beijing Academy of Artificial Intelligence

微信关注
北京智源人工智能研究院