

# 构建强大的深度神经网络：文本方面的调查

王文淇 † §, 王润 † § □, 王丽娜 † § □, IEEE会员, 王志波 §, IEEE会员, 叶傲霜 † §

## 摘要 –

深度神经网络 (DNNs) 在各种任务中取得了显著的成功 (如图像分类、语音识别和自然语言处理 (NLP))。然而，研究人员已经证明基于 DNN 的模型容易受到对抗样本的影响，这些对抗样本通过向合法输入中添加难以察觉的扰动而导致错误预测。最近的研究揭示了文本领域中的对抗样本，这可能有效地规避各种基于DNN的文本分析器，并进一步带来虚假信息传播的威胁。本文对现有的关于生成英文和中文字符对抗性文本的对抗技术研究以及相应的防御方法进行了全面调查。更重要的是，我们希望我们的工作能激发未来研究，开发更强大的基于DNN的文本分析器，以抵御已知和未知的对抗技术。我们根据扰动单元对现有的用于制作对抗性文本的对抗技术进行分类，有助于更好地理解对抗性文本的生成并构建用于防御的强健模型。在介绍文本领域中对抗攻击和防御的分类体系时，我们从不同NLP任务的视角介绍了对抗性技术。最后，我们讨论了文本中对抗攻击和防御的现有挑战，并提出了该新兴而具有挑战性领域的未来研究方向。

索引词 – 对抗攻击与防御、对抗性文本、鲁棒性、深度神经网络、自然语言处理。

## 导言

如今，深度神经网络 (DNNs) 已经展现出在各个领域解决大量具有挑战性问题的强大能力，如计算机视觉[1]、[2]、音频[3]、[4]和自然语言处理 (NLP) [5]、[6]等。由于它们的巨大成功，基于DNN的系统被广泛部署在现实世界中，包括许多安全关键领域[7]–[11]。然而，一系列研究[12]、[13]发现，通过添加几乎察觉不到的扰动而制作的输入可以轻松愚弄DNNs。这些修改后的输入被称为对抗样本，它为基于DNN的系统带来潜在的安全威胁，即使在黑盒情况下，攻击者也无法获得目标系统。例如，图1展示了一种对抗攻击。

† W. Wang, R. Wang, L. Wang 和 A. Ye

就职于教育部航空航天信息安全与可信计算重点实验室

§ W. Wang, R. Wang, L. Wang, Z. Wang 和 A. Ye

就职于武汉大学网络空间安全学院，中国。邮箱：{wangwenqi001, wangrun, lnwang, zbwang, yasfrost}@whu.edu.cn

□ Run Wang 和 Lina Wang 为通讯作者。

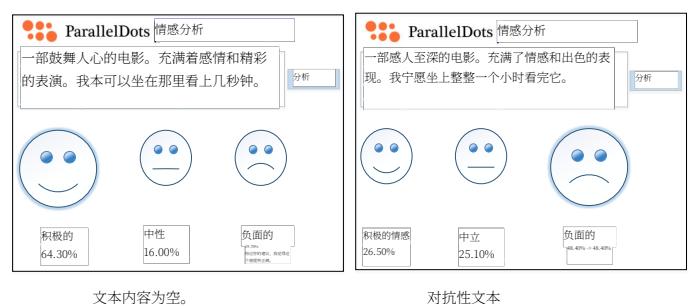


图 1：对流行文本分析系统ParallelDots 进行对抗攻击的实例。ParallelDots 提供了一系列用于各种自然语言处理任务的 API（例如情感分析），在性能上取得了最先进的成果。我们采用基于遗传算法的流行对抗技术来制作对抗性文本，以躲避ParallelDots。我们发现，当原始输入中的单词"inspire"和"wonderful"分别被替换为"touch"和"good"时，文本被以高信心预测为负面。

在名为Parallel-Dots的物理情感分析系统上。在这种情况下，我们无法获得有关系统架构、模型参数和训练数据的任何知识。然而，它未能正确区分对抗性示例，并输出错误结果。为了对抗对抗性示例的威胁，研究人员在攻击和防御方面进行了大量工作，并导致理论和应用技术的显著增加，从图像到文本各种技术变化。在这里，我们关注文本领域中的对抗性示例，而不是广为人知的图像领域。

在自然语言处理中，深度神经网络广泛应用于许多基本任务（例如文本分类、自然语言推理和机器翻译）。不幸的是，这些基于深度神经网络的系统在面对对抗性示例时表现出明显的性能下降。Papernot等人首先发现攻击者可以通过在文本中添加难以察觉的噪音来生成对抗性示例，这将导致分类器产生错误结果。然后，在文本领域战场上开始了一场武装竞赛，导致对抗性示例暴露。

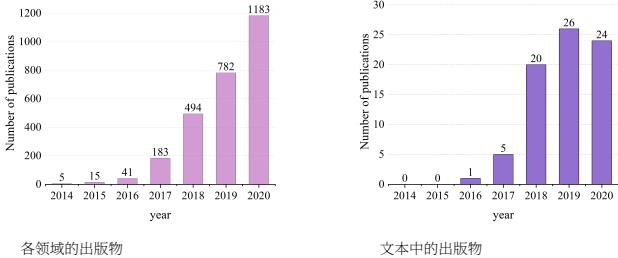


图2：对抗样本的发表刊物。图2(a)展示了对抗样本领域的发表刊物数量，这些数据由Carlini[25]收集，涵盖了广泛的领域，如图像、音频、文本等等。图2(b)代表了对抗文本领域的发表刊物数量。

在这个新兴领域中的研究对抗性攻击主要聚焦于特定的自然语言处理任务[16]–[19]，这将给我们的用户带来潜在的安全问题。例如，在现实世界中，当在线订餐时，用户倾向于在移动应用中搜索附近推荐的餐厅并阅读其产品的评论。服务提供商[20]–[22]会根据发布的评论通过诸如情感分析[23]之类的技术给出建议。然而，这些基于深度神经网络的文本分析器很容易被对抗性样本所欺骗。攻击者可以通过发布对抗性文本来干扰产品评分。更严重的是，攻击者可以恶意通过对抗性文本传播虚假信息以获利并导致消费者损失利润。因此，需要制定有效的防御方法，并为社区开发健壮的模型。

在防御方面，已经提出了一些对抗性样本，以增强基于深度神经网络的文本分析器的鲁棒性。然而，它们显然没有准备好迎接新兴的对抗性样本威胁，因此应继续进行持续努力。图2显示了近年来对抗性样本的出版物，揭示了许多研究正在开发各种对抗性技术，这些技术给防御带来了挑战。目前，对抗性文本检测[24]和模型增强[13]是两种主流思路来对抗对抗性文本的威胁，但它们都存在明显的弱点。例如，对抗性文本检测仅适用于某些对抗性攻击。模型增强如对抗性训练在区分未知对抗性技术生成的对抗性文本方面存在缺陷。总之，应对未知对抗性技术、推广到不同语言以及适用于各种自然语言处理任务是现有防御方法所面临的三大障碍。为了弥补这个明显的差距，迫切需要激励研究者投身于对文本领域中的对抗性攻击和防御的研究。因此，需要进行全面调查以呈现该领域的初步知识并介绍其中的挑战。

在对抗性攻击和防御方面，一些调查专注于图像领域[26]–[31]，但在文本领域中则较少[32]–[34]。这里，我们介绍了这些文本调查中的三项调查，并列出了它们之间的差异。

在2019年3月，Belinkov等人主要关注自然语言处理中机器学习的可解释性。他们仅回顾了一些针对理解这些模型失败的攻击，但他们的工作缺乏对抗对抗性攻击方法的调查。

2020年3月，徐等人[33]系统性地审查了图像、图形和文本领域的尖端算法。对于文本中的对抗性攻击，他们仅根据不同的自然语言处理任务描述了一些方法，但他们并未分析哪种攻击适合该任务，也未比较这些方法之间的相似性和差异。同时，作者们也未关注文本领域的防御。

在2020年4月，Zhang等人[34]

主要比较了图像领域的攻击方法，并描述了对文本中实施对抗攻击的方式。他们将对抗性攻击分为黑盒攻击和白盒攻击，就像在图像领域一样。然而，这种分类方法并不反映如何在自然语言处理中产生对抗性示例。由于文本和图像之间的差异，对抗性示例可以根据文本中的扰动单元划分为字符级攻击、词级攻击、句子级攻击和多级攻击。此外，在自然语言处理中特别设计的防御方法（即拼写检查）并未在他们的防御部分中介绍。

此外，它们都缺乏一些重要的指导原则，比如基于中文和基于英文的对抗样本之间的区别，对对抗样本的可解释性，以及与其他有趣工作的结合（例如，在深度伪造文本中添加对抗扰动以欺骗深度伪造检测器）。

本文回顾了文本领域中对抗样本的研究，旨在通过了解对抗性文本的生成、现有防御方法的弱点与优势，以及不同自然语言处理任务的对抗性技术，来构建稳健的基于深度神经网络的文本分析器。我们工作的进展总结如下。

我们不仅审查文本领域中的对抗性攻击和防御，还包括解释、不可察觉性和认证工作。我们系统性和全面的审查帮助新人理解这一研究领域。

前三次调查仅关注与基于英语的模型相关的作品，并且没有检查评估基于中文模型鲁棒性的努力。我们填补这一空白，分析基于英语和基于中文模型之间对抗性示例的差异。

我们根据对抗性文本中的扰动单元将其分类为字符级、词级、句子级和多级。此外，我们重点关注不同自然语言处理任务的对抗性攻击。我们希望这能激发未来研究者对对抗性文本生成的理解，并进一步开发针对这些自然语言处理任务的通用和有效的防御方法。

我们将对抗性示例与模型分析相结合。