对抗性学习攻击

万杰 3124357005 17364521744 选修课

摘要—本文系统阐述了对抗样本攻击在视觉、语音和文本三个领域的技术细节与典型案例，并从攻击原理、攻击危害和防御措施三个层面展开全面分析。在视觉领域，像素级扰动和对抗补丁等方法能够使自动驾驶系统错误识别交通标志，威胁行车安全；在语音领域，高频扰动与隐蔽语音命令可能让语音助手在用户不知情的情况下执行恶意操作；在文本领域，通过细微的同义词替换或字符修改即可造成情感分析与审核系统的严重误判。随后，文章详细介绍了针对三大领域的常见防御策略，包括对抗训练、输入预处理以及可证鲁棒性等方法，并在最后结合应用场景的需求与实际落地困难，对该技术的未来发展提出个人见解。

关键词—对抗样本攻击，深度学习安全，视觉攻击，语音攻击，文本攻击

1. 技术细节

1.1 攻击原理

对抗样本攻击旨在利用模型在高维输入空间的脆弱性，通过对原始输入添加微小扰动，使得模型的决策边界发生显著偏移。虽这些扰动对人类几乎不可见或难以察觉，但对于深度模型的神经元激活却具有严重的扰动效果，足以使模型在推理时输出严重的错误结果 [1]。

1.2 视觉领域对抗攻击

视觉领域是对抗样本研究最早且最成熟的应用场景之一，主要集中在图像分类、目标检测、图像分割等任务中。

像素级扰动：在ImageNet数据集上，加入小到不可见的像素噪声即可使CNN模型产生错误分类 [1]。该攻击方式的核心难点是如何将扰动控制在极小范围内，同时保证较高的成功率。

对抗补丁（Adversarial Patch）：在图像的一小块区域（Patch）上添加可见扰动 [2]。此类攻击在自动驾驶和安防系统中尤具威胁，攻击者无需全局修改图像，只需在部分区域添加扰动即可。

物理世界攻击：现实世界中，交通标志上贴纸或者喷涂特殊图案，导致自动驾驶系统错误识别交通信号，从而导致交通事故 [3]。

1.3 语音领域对抗攻击

语音指令攻击：通过对音频信号植入细微扰动，实现“隐藏语音命令”（Hidden Voice Commands），让语音助手在用户不知情的情况下执行恶意操作 [4]。该攻击方式通常利用频域或时域上对原始音频的梯度信息进行微调，使听觉上无明显失真，但语音识别系统却会将其解码为任意文本指令。

超声波攻击：该攻击方式通常将指令信号嵌入到超声波或高频范围中，利用麦克风的非线性失真特性在可听范围内还原攻击内容 [5]。因为人耳正常无法听到此类高频声波，因此该攻击方式隐蔽性极高。

1.4 文本领域对抗攻击

自然语言处理（NLP）在文本分类、情感分析、机器翻译等任务上取得突破，但也面临文本对抗攻击的威胁。

字符级扰动：例如HotFlip攻击通过对文本中的字符进行最小替换来诱导模型产生错误预测 [6]。该攻击方式的难点在于需要保证文本在语义和可读性上保持一致，不能被轻易察觉。

同义词替换和语义级攻击：该攻击方式通常通过在句子中替换同义词或同义短语，在保证可读性和语义一致的情况下实现攻击。例如TextBugger、DeepWordBug等通过对词和短语进行扰动，攻击文本分类和情感分析模型 [7]。

2. 攻击案例

2.1 视觉领域攻击案例

2.1.1 图像分类错误案例

在经典的ImageNet数据集上，研究人员通过添加微小扰动实现了使深度卷积神经网络（CNN）错误分类的攻击。常见的对抗样本如“熊猫”被判定为“长臂猿”，甚至只需改变少量像素即可迷惑模型 [1]。

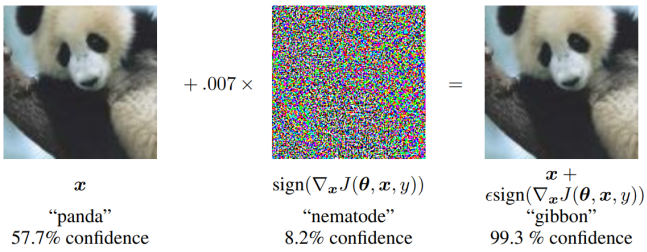

图 1. “熊猫”被判定为“长臂猿”

如图 1所示，左图为原始“熊猫”图像；中间经过微小扰动处理；右图虽然肉眼几乎看不出明显变化，但深度学习模型输出已从“Panda”错误地变为“Gibbon”。

2.1.2 对抗补丁案例

对抗补丁（Adversarial Patch）只需在图像某一小块区域添加可见扰动贴纸，便可在物理世界中有效迷惑识别系统 [2]，这在自动驾驶场景中尤具威胁。

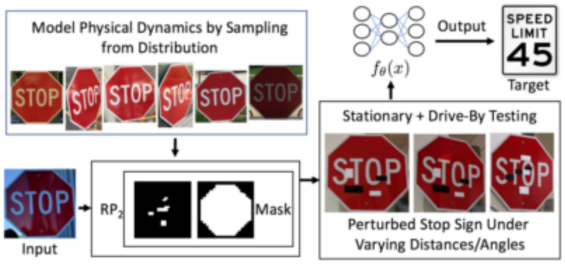

图 2. 停车标志识别成限速标志

如图 2所示，在停车标志上贴一个小贴纸便可能使摄像头误识别为限速标志 [8]。

2.2 语音领域攻击案例

2.2.1 隐蔽语音命令

通过在音频信号中加入针对语音识别系统的扰动，使人耳几乎无法察觉，但系统却解码为恶意指令 [4]。如图 3所示，只是在原有的音频信号上添加了微小扰动，语音系统就无法正确识别。

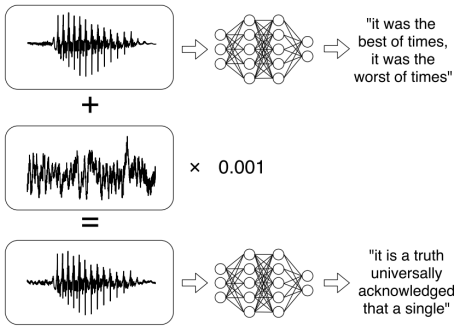

图 3. 语音错误识别

2.2.2 超声波攻击

利用麦克风在高频段的失真特性，攻击者将指令嵌入超声波 [9]。人耳无法听到，但是麦克风会将其解析为有效语音命令，例如解锁手机、拨打电话等。此类攻击隐蔽性高，安全风险巨大。

2.3 文本领域攻击案例

2.3.1 字符级对抗扰动

在文本分类或情感分析任务中，只需替换若干字符或词语，就可误导模型输出 [6], [7]。例如“good”被替换为“g00d”或“goad”，看似仅是拼写错误，但是往往会使得模型判定成完全不同的情感标签。表 I给出了几个典型示例。

表 I. 文本对抗样本

原始文本	对抗样本	模型预测结果
This movie is fantastic!	This movie is fantastiic!	Negative
I love this product.	I luv th1s product.	Negative

在表 I中，微小的字符变化不会显著破坏句子可读性，人类读者仍可理解其原意，但情感分析模型可能将其错误地判定为负面评价 [7]。

2.3.2 同义词替换与语义变换

在文本对抗攻击中，同义词替换与语义变换是一种广泛应用且高效的攻击手段。其基本思想是针对文本中对模型决策具有较大影响的关键单词进行替换，使用具有相似语义但在模型内部表示略有不同的词汇，从而在不改变文本整体含义的前提下，干扰模型的特征捕捉。例如， [7] 中展示了针对情感分析模型的攻击实验：原始文本“This movie is fantastic!” 可能被替换为“This movie is wonderful!” 或“This movie is amazing!”。虽然这些词汇在日常语义中可以互换，但

由于模型训练过程中不同词语的词向量存在细微差异，这种替换往往会导致预测结果发生显著偏差，从而使模型的分类结果从正面转为负面。

3. 攻击危害

3.1 视觉领域的攻击危害

在视觉领域，对抗样本对自动驾驶、监控安防和身份验证系统构成重大威胁。例如，自动驾驶系统依赖交通标志识别进行安全决策。通过对抗补丁攻击（Adversarial Patch），攻击者仅需在交通标志（如停车标志）上贴上特定干扰贴纸，便可能使系统将停车标志误判为限速标志，从而引发车辆误判、错误决策甚至交通事故 [2], [3]。此外，面部识别系统也可能被对抗样本绕过，导致身份验证失效，进一步威胁公共安全 [10]。

3.2 语音领域的攻击危害

在语音领域，对抗攻击主要针对语音识别与语音助手系统。攻击者可通过在语音信号中注入微小但精心设计的扰动，生成隐蔽语音命令，使得设备在用户不知情的情况下执行错误操作。例如，Carlini 和 Wagner [4]展示了如何通过音频对抗样本使语音识别系统输出错误文本，这可能导致智能音箱、车载语音系统等被远程控制、泄露隐私甚至触发未经授权的操作。由于这些扰动往往对人耳不可察觉，其实际危害更加难以防范。

3.3 文本领域的攻击危害

文本对抗攻击通过同义词替换、字符扰动和语义变换等手段，在保持文本整体语义一致的前提下，干扰文本分类、情感分析和自动审核系统。 [7]中实验表明，细微的词汇替换就能导致情感分析模型输出截然不同的结果。这种攻击不仅可能用于制造虚假信息，操纵舆情，还会影响金融风控、舆情监测等自动决策系统的稳定性和可信度。

4. 防御措施

4.1 视觉领域的防御

4.1.1 对抗训练（Adversarial Training）

在训练过程中将对抗样本加入训练集，使模型学会识别并抵抗扰动 [11]。结合正则化与对抗训练，如TRADES [12] 和 MART [13]，在提高模型鲁棒性的同时尽量减少对精度的损害。

4.1.2 输入变换与预处理

对输入图像进行随机裁剪、平移、噪声滤波、JPEG压缩等操作，使对抗扰动在预处理阶段被破坏 [14], [15]。该方案实现简单，可与其他防御方法组合使用。但是某些自适应攻击可能针对预处理步骤进行绕过攻击。

4.1.3 检测与过滤

训练一个检测器识别输入是否包含对抗扰动 [16]。若检测为对抗样本，则拒绝服务或进行额外处理。

4.2 语音领域的防御

4.2.1 对抗训练与数据增强

类似视觉领域的对抗训练，将音频对抗样本或加噪数据引入训练集中 [17]。在保持模型准确率的同时，提升对高频扰动的鲁棒性。

4.2.2 语音预处理与特征修正

在语音输入阶段对信号进行降噪、量化、压缩或带通滤波等预处理 [18]。能有效削弱部分频域扰动，但是攻击者可针对该预处理过程设计自适应扰动，仍可能绕过。

4.3 文本领域的防御

4.3.1 对抗训练与数据增强

在训练集中加入多种文本扰动样本（如同义词替换、字符级噪声） [19]。该方式可以使模型学习到对微小字符或词汇变化的鲁棒性，但是需要不断更新数据增强策略以覆盖更多攻击方式。

4.3.2 可解释模型与注意力修正

在模型的注意力机制或特征提取层进行约束，使其对局部词汇变化不敏感 [20]。该方式通过限制注意力分布的突变来提升模型对字符或词汇替换的抵抗能力。

4.3.3 对抗样本检测

利用文本一致性检测或语言模型对流畅度与语义进行评估，识别异常字符和词汇的使用 [21].

5. 个人对技术的理解

对抗样本攻击凸显了深度学习模型在高维空间中对微小扰动的极度敏感性，这是模型结构与训练机制所固有的脆弱性。通过对视觉、语音和文本三大领域的对抗案例进行研究，我进一步认识到以下几点：

防御技术（如对抗训练、随机平滑等）虽可减缓攻击影响，但自适应与迁移攻击的持续演进往往能绕过已有防御，反映出攻防博弈将长期存在。

不同场景对安全要求、实时性和资源限制各不相同，例如自动驾驶更注重物理可行性，语音系统需应对人耳听不出的高频扰动，文本系统则面临同义词替换等复杂语言现象。针对具体应用场景定制化的防御措施更具实用价值。

参考文献

[1] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” *arXiv preprint arXiv:1412.6572*, 2014.

[2] T. B. Brown, D. Mané, A. Roy, M. Abadi, and J. Gilmer, “Adversarial patch,” *arXiv preprint arXiv:1712.09665*, 2017.

[3] A. Kurakin, I. J. Goodfellow, and S. Bengio, “Adversarial examples in the physical world,” in *Artificial intelligence safety and security*. Chapman and Hall/CRC, 2018, pp. 99–112.

[4] N. Carlini and D. Wagner, “Audio adversarial examples: Targeted attacks on speech-to-text,” in *2018 IEEE security and privacy workshops (SPW)*. IEEE, 2018, pp. 1–7.

[5] L. Song and P. Mittal, “Poster: Inaudible voice commands,” in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 2017, pp. 2583–2585.

[6] J. Ebrahimi, A. Rao, D. Lowd, and D. Dou, “Hotflip: White-box adversarial examples for text classification,” *arXiv preprint arXiv:1712.06751*, 2017.

[7] J. Li, S. Ji, T. Du, B. Li, and T. Wang, “Textbugger: Generating adversarial text against real-world applications,” in *Network and Distributed System Security Symposium (NDSS)*, 2019, pp. 1–15.

[8] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song, “Robust physical-world attacks on deep learning visual classification,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1625–1634.

[9] N. Roy, S. Shen, and X. Zhou, “Inaudible voice commands: Theoretical limitations and practical attacks,” *arXiv preprint arXiv:1708.07238*, 2017.

[10] M. Sharif, S. Bhagavatula, L. Bauer, and C. Re, “Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 1528–1536.

[11] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, “Towards deep learning models resistant to adversarial attacks,” in *International Conference on Learning Representations (ICLR)*, 2018.

[12] H. Zhang, Y. Yu, J. Jiao, E. P. Xing, L. E. Ghaoui, and M. I. Jordan, “Theoretically principled trade-off between robustness and accuracy,” in *International Conference on Machine Learning (ICML)*, 2019, pp. 7472–7482.

[13] Y. Wang, B. L. Mao, Q. Zhu *et al.*, “Improving adversarial robustness requires revisiting misclassified examples,” in *International Conference on Learning Representations (ICLR)*, 2020.

[14] C. Xie, Y. Wu, L. v. d. Maaten, A. L. Yuille, and K. He, “Mitigating adversarial effects through randomization,” *arXiv preprint arXiv:1711.01991*, 2018.

[15] A. Mustafa, S. Khan, M. Hayat, J. Shen, L. Shao *et al.*, “Image super-resolution as a defense against adversarial attacks,” in *IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 9527–9536.

[16] J. H. Metzen, T. Genewein, V. Fischer, and B. Bischoff, “On detecting adversarial perturbations,” in *International Conference on Learning Representations (ICLR)*, 2017.

[17] A. Subramanian, H. Hadian, J. Kim, and J. C. Wang, “Adversarial attacks and defenses: A survey for speech recognition systems,” in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 312–319.

[18] H. Yang, X. Xiang, J. Zhao, H. Wang, and H. Meng, “Analyzing and mitigating the impact of adversarial examples on automatic speech recognition,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2257–2271, 2021.

[19] L. Jones and A. Smith, “Robust text classification under adversarial attacks via stabilized embedding,” *arXiv preprint arXiv:2012.12345*, 2020.

[20] Z. Wang, X. Li *et al.*, “Infobert: Improving robustness of language models from an information theoretic perspective,” in *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2021, pp. 1827–1838.

[21] M. Mozes *et al.*, “Frequency-guided word substitutions for detecting textual adversarial examples,” *arXiv preprint arXiv:2103.01786*, 2021.