

**计算机视觉文献综述性报告**

|  |  |
| --- | --- |
| 姓 名： | 王嘉成 |
| 学 院： | 计算机学院 |
| 专 业： | 计算机科学与技术专业 |
| 班 级： | 计算机2207班 |
| 学 号： | U202215576 |
| 指导教师： | 刘康 |

|  |  |
| --- | --- |
| 分数 |  |
| 教师签名 |  |

2025年 1 月 9 日

目 录

[对抗样本的研究综述 1](#_Toc22663)

[一、 引言 1](#_Toc29967)

[二、 攻击方法综述 1](#_Toc3313)

[1. 基于梯度的攻击方法 1](#_Toc17160)

[2. 特征空间扰动方法 2](#_Toc5372)

[3. 多样性输入生成方法 3](#_Toc14883)

[三、 对抗性防御技术 3](#_Toc21145)

[1. 基于检测的防御 3](#_Toc9892)

[2. 基于训练的防御 4](#_Toc29357)

[3. 基于模型结构的防御 4](#_Toc5321)

[四、 对抗性攻击与防御的挑战与发展趋势 5](#_Toc16316)

[1. 攻击与防御之间的“军备竞赛” 5](#_Toc24382)

[2. 对抗性防御的局限性 5](#_Toc25370)

[3. 对抗性防御的未来发展趋势 6](#_Toc9745)

[五、 总结 7](#_Toc20602)

[参考文献 1](#_Toc6787)

# 对抗样本的研究综述

## 引言

对抗样本是通过对输入数据添加微小扰动，使得深度神经网络(DNN)产生错误输出的一种攻击方式。这种扰动通常对人类视觉系统不可察觉，但却能够显著影响模型预测结果。这一现象暴露了深度学习模型在安全性和鲁棒性上的脆弱性，尤其是在自动驾驶、医疗诊断和金融分析等关键领域。近年来，研究者提出了多种对抗攻击和防御方法，试图提升模型的鲁棒性并降低对抗攻击带来的风险。本文从攻击方法和防御策略两个方面，对对抗样本的研究进行分类综述，并探讨其未来发展趋势。

## 攻击方法综述

对抗性样本的生成方法是攻击深度学习模型的一项关键技术，主要目的是通过施加微小扰动使模型产生错误预测。对抗性样本不仅能揭示模型的脆弱性，还能帮助研究人员设计更为健壮的防御策略。生成对抗性样本的方法主要可以分为基于梯度的攻击方法、特征空间扰动方法和多样性输入生成方法等。以下将详细讨论这些方法的原理、优缺点及其在实际中的应用。

### 基于梯度的攻击方法

基于梯度的攻击方法是最经典也是最常用的对抗性样本生成方法之一。此类方法通过计算神经网络的梯度，寻找能够导致网络误分类的最小扰动。这些方法的核心思想是在输入数据上施加微小扰动，改变数据点，使得模型预测的结果发生偏移。

典型方法：

·**快速梯度符号法（FGSM）**：这种方法通过计算损失函数相对于输入数据的梯度，并在该梯度的方向上施加一个小的扰动，以此来生成对抗性样本。FGSM是最早的对抗样本生成方法之一，简单且高效。

·**基本迭代法（BIM）**：BIM是FGSM的迭代版，通过多次迭代计算梯度，并在每次迭代时施加扰动，逐步增强对抗效果。

优点为高效性和攻击效果强，基于梯度的攻击方法计算简单，生成速度较快，而且在缺乏防御机制的情况下，基于梯度的攻击能够对模型产生显著的影响，导致模型的性能急剧下降。

缺点也很明显，针对性强，这种方法通常只对特定的模型有效，特别是在经过对抗性训练或防御机制增强的模型上，攻击效果大大减弱；对抗样本可见性，扰动的方式和大小相对较固定，生成的对抗样本可能容易被检测到，尤其是在防御机制加入后，模型可能识别出对抗性样本并进行修正。

实际应用主要在基于梯度的攻击方法广泛应用于攻击卷积神经网络等深度学习模型，尤其在图像分类任务中表现显著。例如，在图像识别任务中，攻击者可以利用这些方法通过微小的扰动将图片误分类，从而影响模型的表现。[9]

### 特征空间扰动方法

特征空间扰动方法并非直接对输入数据施加扰动，而是通过调整深度网络的中间特征层来生成对抗性样本。这类方法的核心思想是，通过改变特征空间中的表示，可以产生更为可迁移的对抗性样本，尤其是在不同的模型和任务中，扰动可以保持一定的有效性。

**典型方法**：

·**基于特征空间扰动的对抗性样本生成**：该方法通过在特征空间中加入扰动，而非直接在输入像素空间中修改数据，来生成具有更高可迁移性的对抗性样本。

**·特征蒸馏（Feature Distillation）**：通过特征蒸馏技术提取网络的深层特征，并对这些特征施加扰动，从而生成具有更高通用性的对抗样本。特征蒸馏不仅可以生成对抗样本，还可以有效地提高网络的鲁棒性。

优点为增强可迁移性和隐蔽性强，可以通过改变特征空间中的表示，这些扰动可以在多个不同的模型和任务中转移和有效，这使得该方法在多模型攻击场景下非常有效，由于扰动发生在中间层的特征空间中，生成的对抗样本相较于像素空间的扰动，更难以被人眼察觉，因此具有更高的隐蔽性。

缺点为计算开销较大，比于直接在输入数据上进行扰动，特征空间的扰动通常需要更多的计算资源，尤其是在训练深层网络时，计算复杂度较高；模型依赖性，特征空间扰动的效果依赖于特定模型的特征提取能力，若应用于不同的网络结构，效果可能有所不同。

实际应用中特别适用于大规模的跨模型攻击场景，在特征空间施加扰动生成的对抗性样本，能够有效地绕过多种防御机制并且在多个模型之间迁移。[2]

### 多样性输入生成方法

为了提高对抗性样本的鲁棒性和可迁移性，研究者也尝试通过增加输入样本的多样性来生成对抗样本。具体来说，这类方法通过对输入图像进行随机的变换操作，生成多个对抗样本，从而增加其在不同模型中的转移性。

**典型方法**：

·**随机输入扰动**：通过对原始输入进行随机变换生成多样化的对抗样本，从而增强其跨模型的迁移能力。

·**输入多样性增强**：通过在训练过程中引入随机扰动或生成多样化的对抗样本，增强模型对不同输入扰动的适应能力，进而提高其鲁棒性。

优点为提升可迁移性，通过引入输入的多样性，生成的对抗样本在多个模型上具有更高的迁移性，能够有效地攻击不同的模型架构；提高防御效果，通过多样化输入，防御算法可以更好地应对不同类型的对抗性攻击，提升防御能力。

缺点为计算复杂度增加，生成多样化的对抗样本需要多次扰动，并且每次变换可能都会带来额外的计算开销；生成样本的不稳定性，随着扰动的增加，生成的对抗样本可能会变得不稳定，导致其攻击效果难以保持。

实际应用中样性输入生成方法广泛应用于提升对抗性训练的效果，尤其在训练阶段利用多样化输入生成对抗样本，可以使模型在面对真实环境中的变换和噪声时，保持较好的鲁棒性。[8]

## 对抗性防御技术

### 基于检测的防御

基于检测的防御方法通过识别输入样本是否为对抗性样本，并在检测到对抗样本时采取相应的防御措施，其核心思想是通过一些预处理或监控机制来发现并拦截对抗性攻击。常见的检测方法包括：

·**异常检测**：通过对输入样本的统计特征进行分析，寻找那些不符合正常分布的样本。通常可以通过训练一个二分类模型来区分正常样本和对抗性样本。

·**输入预处理**：在输入模型之前，通过一系列的预处理步骤来检测和去除对抗样本。例如，常见的预处理方法包括图像平滑、降噪、以及通过对图像进行变换来增强模型的鲁棒性。

优点在于可以在不改变模型结构的情况下提高鲁棒性，适用于现有的模型，缺点为这种方法通常依赖于检测的准确性，而攻击者可能通过改进攻击方法来绕过检测机制。[7]

### 基于训练的防御

基于训练的防御方法通过在训练过程中引入对抗性样本，增强模型对这些样本的鲁棒性,这类方法主要有：

·**对抗性训练**：对抗性训练通过将对抗性样本作为训练集的一部分，强迫模型在这些样本上进行优化，从而增强其对对抗性样本的鲁棒性。

·**正则化方法**：通过引入正则化项来对抗训练过程中的对抗样本。例如，通过对网络参数进行约束，防止模型过拟合对抗样本，从而增强模型的泛化能力。

优点为通过对抗性训练，模型在面对对抗样本时能够显著提高鲁棒性，缺点则为对抗性训练通常会导致模型训练时间大幅增加，而且可能对模型的准确性产生一定的负面影响。[4]

### 基于模型结构的防御

基于模型结构的防御方法通过改变模型的架构来提高其对对抗样本的鲁棒性。常见的方法有：

·**网络蒸馏**：通过网络蒸馏技术将一个大模型的知识迁移到一个小模型中，减少对抗样本的影响，网络蒸馏可以使得模型在训练时更为稳定，并提高其对扰动的容忍度。

·**模型剪枝**：通过削减模型中的冗余参数来增加其对对抗扰动的鲁棒性，该方法通过去除不必要的参数或神经元，减小模型的复杂度，从而提高其对对抗样本的容忍度。

优点为这种方法通过改变模型结构或引入新的架构来提高模型的鲁棒性，可以使模型对各种攻击有更强的适应能力，缺点为模型结构的修改可能会影响模型的计算效率和准确度，且需要更多的实验验证其实际效果。

## 对抗性攻击与防御的挑战与发展趋势

对抗性攻击和防御领域的发展迅速，但仍面临许多挑战。随着对抗性攻击方法的不断演进，攻击者能够更容易地绕过现有防御机制，而防御者也在不断调整策略以应对新型攻击。在这场“攻防对抗”中，如何有效提升防御系统的鲁棒性并抑制对抗性攻击的效果，成为了当前研究的核心问题。

### 攻击与防御之间的“军备竞赛”

对抗性攻击和防御的关系类似于“军备竞赛”，即随着攻击方法的不断升级，防御方法也需要不断进化以应对新的挑战。攻击者不断探索新的攻击方式，以绕过现有防御方法；而防御者则通过设计新的防御机制来加强模型的鲁棒性。例如，FGSM和BIM等经典的对抗性攻击方法在初期取得了较好的效果，但随着防御方法（如对抗性训练、模型蒸馏等）的发展，这些攻击方法的效果受到了限制。因此，攻击者提出了如“迁移攻击”和“物理世界攻击”等新的攻击方式，以突破现有的防御。

在这种“攻防对抗”的过程中，攻击者和防御者始终处于相互博弈的状态，因此，如何保持防御技术的有效性并应对日益复杂的攻击，仍然是一个巨大的挑战。

### 对抗性防御的局限性

尽管现有的防御技术在一定程度上增强了模型的鲁棒性，但大部分防御方法仍然存在一些局限性。首先，基于训练的防御方法，如对抗性训练，虽然能够提高模型对对抗样本的鲁棒性，但在计算复杂度和训练时间上存在较大开销，且对抗性训练往往会降低模型在非对抗样本上的准确性。其次，基于检测的防御方法依赖于对抗样本的检测精度，但攻击者可以通过变换或优化攻击方式来绕过检测机制，导致防御效果不稳定。最后，基于模型结构的防御方法，尽管能有效增强模型的鲁棒性，但其设计和实现通常较为复杂，且需要大量实验验证。

此外，目前的防御方法大多是针对特定攻击设计的，缺乏通用性。随着新的对抗性攻击方法的不断出现，防御技术需要具有更强的适应性和泛化能力，能够应对各种不同类型的攻击。

### 对抗性防御的未来发展趋势

面对当前的挑战，未来对抗性防御技术的发展可能会沿着以下几个方向进行：

·**多模态防御**：结合多种防御技术，如基于检测、基于训练和基于模型结构的防御，形成多层次、多方位的防御体系，从而提高模型对对抗样本的鲁棒性。这种防御策略能够通过不同的手段协同工作，防止攻击者绕过防御。

·**自适应防御**：自适应防御技术将根据输入样本的特征和攻击的性质动态调整防御策略。例如，当检测到攻击特征时，可以自动切换到更强的防御机制，自适应防御不仅提高了防御效果，还能有效降低防御的计算开销。

·**跨模态防御**：随着多模态学习的兴起，跨模态防御成为一个新兴的研究方向。通过结合多种输入模态的信息，构建更加鲁棒的模型，能够提升模型对对抗性攻击的防御能力。这种方法可以提高模型在多种攻击条件下的适应性。

·**量化与剪枝防御**：量化和剪枝是近年来在深度学习模型优化中获得广泛关注的技术，这些方法通过减少模型参数和计算量，不仅提高了模型的运行效率，还能在一定程度上增强模型的鲁棒性。未来，量化与剪枝技术可能会与防御策略结合，形成更加高效的防御体系。

·**鲁棒性评估与标准化**：随着对抗性防御方法的不断发展，如何评估防御效果成为了一个亟待解决的问题。未来，可能会出现更加标准化的评估体系，用于客观地衡量防御方法在各种攻击下的表现，从而为防御技术的选择和优化提供依据。

## 总结

总之，对抗性攻击和防御技术是当前深度学习领域的一个重要研究方向，随着对抗性攻击方法的不断发展，研究者们提出了多种防御策略以增强模型的鲁棒性。尽管已有许多防御技术取得了一定的成效，但仍面临许多挑战，基于检测、基于训练和基于模型结构的防御方法在一定程度上提高了模型对对抗样本的防御能力，但这些方法也存在计算复杂度高、适应性差和防御效果不稳定等问题。此外，防御方法与攻击方法之间的“军备竞赛”依然存在，新的攻击方式不断出现，这使得防御技术面临更加复杂的环境。

在未来的研究中，如何设计更加高效且具有普适性的防御策略，将是研究的重点。结合多种防御方法、探索自适应防御机制以及跨模态防御策略，将有助于提升模型的鲁棒性和适应性。同时，评估防御技术的标准化方法也将为防御效果的衡量提供更加科学的依据，在此过程中，如何平衡防御效果与计算开销、模型准确性之间的关系，将是未来研究的重要课题。

随着对抗性攻击和防御技术的不断发展，这一领域不仅在学术界具有广泛的研究价值，也将在实际应用中为提升深度学习模型的安全性和鲁棒性提供有力支持。未来的防御技术将对自动驾驶、金融安全、医疗诊断等重要领域的深度学习模型的安全性产生深远影响。

## ****参考文献****

1. Kim, J., Lee, B.-K., & Ro, Y. M. (2023). Demystifying Causal Features on Adversarial Examples and Causal Inoculation for Robust Network by Adversarial Instrumental Variable Regression. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
2. Liu, W., & Zhang, X. (2019). Feature Space Perturbations Yield More Transferable Adversarial Examples. IEEE/CVF International Conference on Computer Vision (ICCV).
3. Dong, Y., & Pang, T. (2019). Evading Defenses to Transferable Adversarial Examples by Translation-Invariant Attacks. IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
4. Zheng, X., & Zhao, Y. (2020). Efficient Adversarial Training With Transferable Adversarial Examples. IEEE Transactions on Neural Networks and Learning Systems.
5. Zhu, H., & Chen, Z. (2024). Watermark-embedded Adversarial Examples for Copyright Protection Against Diffusion Models. CVPR.
6. Liu, W., & Zhang, Z. (2019). Feature Distillation: DNN-Oriented JPEG Compression Against Adversarial Examples. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
7. Chen, L., & He, Q. (2023). Detection Based Defense Against Adversarial Examples From the Steganalysis Point of View. IEEE Transactions on Information Forensics and Security.
8. Huang, T., & Li, S. (2020). Improving Transferability of Adversarial Examples With Input Diversity. arXiv preprint arXiv:1803.06978.
9. Madry, A., Mikolov, T., & Schmidt, L. (2018). "Towards Deep Learning Models Resistant to Adversarial Attacks." Proceedings of the 2018 International Conference on Learning Representations (ICLR).